

Bioinformatics expertise has lagged behind sequencing technology. Groups still do not agree on a standard way to process the information. Still this technology improves rapidly, and recently a group published 24-hour genome sequencing for intended us in clinical decision making [19]. Soon it may be a reality for physicians to utilize genomic information, whether about drug susceptibility, or prognosis, to guide medical care. Here we review the methods to asses genetic changes. We discuss issues that present with each method.

## 2 GENOME ANALYSIS

### 2.1 Chromosome Analysis

Historical mainstays to asses changes in the human genome include a method known as a karyotype analysis. A karyotype visualizes the 23 chromosomes that contain our genetic information. Aneuploidy is duplication of a chromosome. Trisomy 21 is a well known syndrome characterized by duplication of the 21st chromosome. Duplication or deletion of all other chromosomes is not compatible with life. However, portions of chromosomes can be duplicated or deleted, giving rise to well known syndromes. Karyotype analysis is capable of visualizing large deletions and duplication in chromosomes, generally greater than 10Mb. Chromosome analysis has been largely surpassed by newer technologies. Given established use and accessibility, it may have a clinical role in rapidly confirming a suspected aneuploidy.

### 2.2 Flourescent In Situ Hybridization

Fluorescent In Situ Hybridization utilizes fluorescent labeled probes to identify portions of DNA which match the probe sites. In this way the chromosomal material can be visualized. Fluorescent In Situ Hybridization can identify chromosomal duplication and deletions up to 2MB. This is helpful, to identify large duplication's and deletions leading to disease. However we know even single nucleate changes lead to disease. Therefore Fluorescent In Situ Hybridization has been replaced by other technologies [2].

### 2.3 Genome Wide Association Studies

Historically research has focused on aneuploidy and syndromes representing large duplication or deletion of genetic material, or on single gene mutations leading to disease. However pathogenesis likely involves multiple common and rare single nucleotide variants (Single nucleotide variation) in parallel leading to most disease. Genome Wide Association Studies emerged to study common variants on large scale, and studies have showed multiple susceptibility loci8. However, Genome Wide Association Studies failed to identify all forms of genetic disease [24].

### 2.4 Copy Number Variation

A large part of the human genome consists of repetitive sequence, including both long and short repeated segments. There are distinct regions that vary in the number of repeats between individuals, and this variation leads to phenotypic differences between these individuals. This variation is referred to as copy number variation (Copy Number Variation). It is thought that up to 10% of the genome consists of Copy Number Variation. Most Copy Number Variation is inherited but it can also occur de-novo. Copy Number

Variation is increasingly understood as contributing to disease, where varying amounts, or doses, of a particular gene and therefore protein lead to disease [32].

### 2.5 Chromosomal Microarray

Chromosomal microarray is the baseline genetic testing for individuals with disease. Chromosomal Microarray is a technology that detects the presence or absence of patient DNA by measuring hybridization of patient sample to small segments of DNA attached to a surface. Chromosomal Microarray detects deletions and duplications of chromosomal material much smaller than FISH and karyotype. As technology improves, Chromosomal Microarray is able to detect increasingly small changes down to, but excluding, Single nucleotide variation. As many common diseases are due to Single nucleotide variation, sequencing is often necessary. [26]

### 2.6 Sanger Sequencing

In 1977 a paper was published entitled “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. This technique, now known as Sanger sequencing, revolutionized molecular biology. Using termination of sequence and dye detection, it provided a fast and easy way to determine the DNA sequence of living organisms. It is still extensively utilized. Many newer technologies have been developed and are known as “next generation sequencing.” [20]

### 2.7 Next Generation Sequencing

Next Generation Sequencing refers to a variety of technologies and a number of different methods for high throughput sequencing of DNA samples [16]. The technology utilizes massive number of parallel sequencers to copy short fragments of DNA and assemble transcripts utilizing big data and bioinformatics techniques. According to the illumina website “With its unprecedented throughput, scalability, and speed, next-generation sequencing enables researchers to study biological systems at a level never before possible.”

### 2.8 Targeted Gene or Gene Panel Sequencing

Disease is often due to Single nucleotide variation necessitating sequencing for diagnosis. Targeted sequencing is commercially available to detect Single nucleotide variation in a specific gene, or an entire panel of genes, often depending on the disease. Gene panels are available for particular syndromes. Commercial panels utilize both traditional sanger sequencing and NGS technology. Targeted sequencing often provides better coverage of specific genes than does Whole Exome Sequencing. This targeted sequencing circumvents the significant burden of analyzing thousands of variants of unknown significance, a problem inherent to Whole Exome Sequencing, but misses variants in genes outside of the panel, or in novel genes.

### 3 NEXT GENERATION SEQUENCING

#### 3.1 Whole Exome Sequencing

With advancements in technology, exome sequencing is approaching the affordability and efficiency of targeted gene panel sequencing. Whole exome sequencing involves sequencing the entire coding region, or exome, of the genome. This consists of around 20,000 genes and over 30 million nucleotides. The exome, though massive, consists of only 1% of the total genomic DNA. Most genetic diseases involve alteration of this coding exome. Sequencing only 1% of the genomic material is a fraction of the time, cost, and burden of analysis, compared with Whole Genome Sequencing. Due to errors in Whole Exome Sequencing, a portion, (up to 1%), of the coding exome is missed. Coverage varies by gene and by region, with particular genes of interest, such as the HRAS gene implicated in Costello Syndrome, difficult to capture by Whole Exome Sequencing at all. Copy Number Variation, insertions, and deletions are also difficult to detect. Targeted sequencing is often advantageous, but Whole Exome Sequencing is improving and is increasingly accessible to clinicians [30].

#### 3.2 Whole Genome Sequencing

Despite the massive amount of information produced in Whole Exome Sequencing, it represents only 1% of the total genome. Transcription enhancers and promoters, often involved in disease pathogenesis, are outside of the exon and missed by Whole Exome Sequencing. In addition, Whole Genome Sequencing better captures Copy Number Variation, insertions and deletions, frequently involved in disease. Whole Genome Sequencing adds significantly to expense, data storage, analysis, and the burden of determining variant significance. For this reason Whole Genome Sequencing is predominantly used in the research setting, but this is changing. In 2012 a group used rapid Whole Genome Sequencing in the newborn ICU to identifying disease causing pathologic variants. The process, from sample collection to automated bioinformatics analysis, was complete within in 48 hours. This rapid turnaround was intended as a model for utilization of Whole Genome Sequencing in clinical decision making. As technology improves Whole Genome Sequencing will likely become a useful clinical tool [13].

#### 3.3 Variants of Unknown Significance(VUS)

Whole Exome Sequencing produces tens of thousands of variants, and Whole Genome Sequencing exponentially more. Another major hurdle is determining significance. Each variant must be assessed for disease pathogenesis, distinguishing it from a previously unreported polymorphism. Variants can be filtered for pathogenic nature based on conservation across populations and location in a protein. It is often necessary to obtain parents samples and perform sequencing on patient-parent trios to determine novelty. When a novel variant is identified, ideally biologic mechanism is investigated through animal and cell culture models. Genetic variation can now be introduced into animal and cell culture models with greater ease and efficiency utilizing the CRISPR-Cas9 system. Variants are often damaging only in conjunction with other variants. In some cases it is impossible to narrow down a single candidate when a

disease with incomplete penetrance and variable expressivity affects a small family. Efforts are ongoing to improve and streamline variant analysis for clinical utilization [14].

### 4 BEYOND DNA

Initial estimates placed the number of genes at  $\approx 100,000$  [1]. Looking at the massive amount of diversity and the billions of unique human beings on this earth, this was an appropriate estimate. The current number is estimated somewhere around 20,000. The question is what accounts for the rest of phenotypic diversity and disease. The picture of development is complex with networks of genes turned on and off at different locations and timepoints. Regulation of this process occurs to some extent outside of the coding region, through promoters and enhancers, epigenetic alterations, splicing variation, and noncoding RNA. Altered noncoding sequence is increasingly implicated in disease. The human genome project utilized whole exome sequencing. The exome, though massive, consists of only 1% of the total genomic DNA. Many genetic diseases involve alteration of this coding exome but we are discovering that many diseases are due to problems outside of this coding region. Whole genome captures this noncoding region, although with far greater cost, burden of analysis, etc. We have also come to realize that splicing and other post transactional regulation introduces much diversity. We have the technology to sequence the entire RNA transcriptome and the proteome as well. This produces a data set which dwarfs the genome and genomic DNA sequence information. These technologies are currently only utilized in the research setting. Despite our advanced technology, we have very little idea of how to interpret the data in a clinical setting. Again the bioinformatics expertise lags behind. There is amazing potential to advance knowledge and study human disease and a tremendous amount of big data analytics along the way.

#### 4.1 RNA Sequencing

Splicing variation leads to multiple different proteins resulting from a single gene due to differential splicing during transcription. Non-coding RNA also influences expression and modifies proteins after translation. Technologies to examine the elements, include Whole Genome Sequencing to measure DNA outside of the exome, RNA sequencing to measure the splice variants and the transcriptome, and ChIPSeq to measure DNA methylation. These noncoding regulatory elements have important clinical implications, and need further exploration. [25]

#### 4.2 Epigenetic Sequencing

Mutations in transcription factors are well established in pathogenesis and regulation is often through enhancers and promoters outside of the coding sequence. The term epigenetics refers to alterations outside of, or on top of, the genetic material or DNA, that influence phenotype. Common epigenetic factors include DNA methylation, where methylation of DNA bases represses DNA expression, and also histone modification, where the degree to which DNA is wrapped around histones influences its expression. Epigenetic factors are heritable, and also influenced by the environment. [11]

### 4.3 Proteome

Ultimately DNA codes for RNA and RNA is translated into proteins. Proteins are the building blocks of all living things. The proteome is the term for the entire protein content of an organism. New technologies allow us to measure the proteome. The proteome is generally measured through tandem mass spectroscopy or finger-printing. Tandem mass spectroscopy breaks proteins into smaller portions and measures a signature and electrophoresis techniques involve separating proteins on a gel and measuring their fingerprint. These techniques require sophisticated chemistry and data analysis techniques and produce massive datasets. [8]

### 4.4 Metabolome

The Metabolome involves the entire set of small molecules within an organism. Analyzing the metabolome involves measuring every amino acid, organic acid, vitamin and mineral in a cell, tissue or organism. Measurement is usually by mass spectroscopy or nuclear magnetic resonance spectroscopy. requires extensive data analysis.

## 5 OTHER GENOMICS TOOLS

Sequencing technology is not the only factor revolutionizing personalized medicine. There is a separate and equally exciting revolution in cell culture technology integral to personalized medicine. All of these technologies rely on genomics measurements that produce massive datasets and rely on Big Data for analysis.

### 5.1 Model Systems

The optimal diagnosis and treatment of pediatric disease requires an understanding of physiology and pathophysiology. Throughout medical research history animal and cell culture models have been critical to this process. Mouse models, in particular, are extensively utilized because they are relatively convenient, and similar to humans at the chemical, molecular, cellular, and some anatomic levels. Furthermore, the use of transgenic mice allows for genetic manipulation to help elucidate molecular mechanisms. However, given that mice and humans diverged millions of years ago, there are critical physiological differences between the two species. Human diseases often lack a mice ortholog. The equivalent disease in mice may be fatal or benign, and we cannot model some high level human organ functions or late onset diseases. Even non-human primates, despite being our closest ancestors, have important phenotypic differences. For example, because of these differences, it is particularly difficult to develop animal models for neurodegenerative or neurodevelopmental disorders. Differences in mouse disease morphogenesis have led difficulty modeling human congenital heart disease. These limitations drive the need for human cell, tissue, and organ systems models. Many human diseases involve terminally differentiated cell types, such as neurons and cardiomyocytes. These cell types are nearly impossible to sample, culture, and maintain. Even after generating primary cell lines from diseased tissues, ability to derive meaningful conclusions is often hampered by inconsistent replicability, dedifferentiation, and variability due to culture conditions. In this light, tissues derived from human induced pluripotent stem cells (h induced pluripotent stem cellss) has the potential to overcome many inherent limitations of animal and cell culture models

and provide an unprecedented new paradigm to model human diseases.

### 5.2 Pluripotent Stem Cells

During human embryogenesis, the ovum and spermatozoa fuse at fertilization, begin to divide, and differentiate into all cell lineages and tissue types in the human body. During development, these cells lose their pluripotency as they terminally differentiate into specific cell types. Embryonic stem cells (ESC) were first isolated from the blastocyst of developing mouse embryos in 1981, and from human embryos in 1998 [17]. These cells have the remarkable ability to retain pluripotency. The ESC discovery generated great excitement over their potential applicability in human disease modeling and regenerative therapies. However, limitations and controversies soon emerged. The isolation of ESCs from human embryos is ethically controversial. Disease models utilizing ESC are limited to diseases identified through preimplantation genetic diagnosis. Genome editing ECSs provides an opportunity to generate particular mutations of interest, but technique remains largely limited to monogenic diseases. In this light, recent breakthroughs in induced pluripotent stem cell ( induced pluripotent stem cells) technology circumvent many of these drawbacks.

### 5.3 Induced Pluripotent Stem Cells

In 2006, Shinya Yamanaka identified four transcription factors, (OCT4, SOX2, KLF4, and c-MYC), that were capable for reprogramming somatic mouse cells into a pluripotent state [22]. This extraordinary feat was recapitulated one year later in human cells. These induced pluripotent stem cells ( induced pluripotent stem cellss) behave like ESCs with capability to differentiate to most other cell types, and circumvent the ethical controversy and sample limitations. As opposed to human embryos, induced pluripotent stem cellss can be generated from readily accessible tissue samples, such as peripheral blood mononucleated cells (PBMCs). Patient samples can be reprogrammed to induced pluripotent stem cellss, serving as an autologous, continuously renewing supply of pluripotent cells. This has resulted in the dramatic expansion of the stem cell field, with development and improvements in reprogramming protocols, and directed cellular differentiation. Patient-specific induced pluripotent stem cellss can be generated from wide variety of patient samples, including PBMCs from blood samples, to dermal fibroblasts from punch biopsies, and epithelial cells from urine samples. induced pluripotent stem cellss can then be differentiated to most other cell types including cardiomyocytes, neurons, and hepatocytes. Because the lines are patient-specific, they are expected to recapitulate features of many disease phenotypes, whether due to simple monogenic mutations or complex polygenic disease susceptibilities. The patient-specific induced pluripotent stem cellss hold potential for disease modeling, predicting drug response, assessing environmental triggers of diseases, and regenerative tissue engineering. Thus, they provide great potential for research and clinical applications in personalized medicine.

## 5.4 Gene Editing induced pluripotent stem cellss

Mouse models allow genetic alteration using transgenesis and gene knock-outs. Measuring the resulting phenotype is extremely valuable in the study of genetics and development. induced pluripotent stem cellss allow us to utilize these same genetic approaches using human cell lines. The past decade has seen tremendous advances in gene editing technology, including ZFNs (zinc finger nucleases), TALENs (transcription activator like effector nucleases), and CRISPRf?Cas9 (clustered regularly interspaced short palindromic repeat) [21] [10]. The common mechanism of these genomic editing approaches is that they create double stranded breaks (DSBs) at desired locations in the genome, which then can be repaired by either nonhomologous end-joining (NHEJ) that can result in insertion/deletions (indels) or homology directed repair (HDR), which results in precise gene modifications. Of these, the CRISPR-Cas9 technology, which appropriates the prokaryote defense mechanism, has quickly become dominant due to ease with which it can be adapted to precisely edit virtually any region in the host genome. Genome editing, coupled with the induced pluripotent stem cells technology, allow us to study disease mechanism like never before. These technologies allow us to precisely correct mutations, and insert reporters under the endogenous regulatory control. They have also been used to demonstrate feasibility of genomic editing as a therapeutic modality. Recently, a group corrected a pathogenic mutation in preimplantation human embryos, demonstrating the feasibility of gene correction therapy. While still a long way from clinical applications, many disease phenotypes have been corrected in cell culture. These studies show the potential of these powerful technologies for disease modeling, and for therapeutic genome engineering.

## 5.5 Organoid Models

Sometimes a simple, two-dimensional induced pluripotent stem cells-derived tissue culture model cannot fully recapitulate complex organ systems involving three dimensional (3D) architecture; such cases necessitate organoid modeling. In vitro organogenesis, the exciting new frontier in in vitro disease modeling, aims to organize induced pluripotent stem cellss into 3D structures that better recapitulate in vivo physiology. Previous attempts at organoid modeling utilized primary tissue cells, but primary cells are difficult to obtain and often fails to propagate in vitro. In principle, induced pluripotent stem cellss are an ideal cell source to make tissue organoids. The most comprehensive organoid model to date involves a fully vascularized and functional human liver. A 3D gastric organoid was created that progresses through developmental stages adopts similar architecture to the stomach. This organoid provided valuable insights into the gut development, as well as H. Pylori infection. Human induced pluripotent stem cellss were grown also on rat intestinal matrix, to engineer a humanized intestinal graft for nutrient absorption in patients with short bowel syndrome. The established protocol for generating 3D cerebral organoids from induced pluripotent stem cellss, replicates brain developmental stages. The organoid reproduces a variety of brain structures, including the cerebral cortex, ventral telencephalon, choroid plexus

and retina. Manipulating specific developmental signaling pathways in ventral-anterior foregut spheroids recently generated an induced pluripotent stem cells-based human lung model. Lastly, an induced pluripotent stem cells-based human kidney organoid model was recently developed displaying glomerulus-like structures and renal tubules. Future in vitro organogenesis effort must address the need for chemically defined synthetic extracellular matrices, and incorporation of support cell types such as interspersed neurons, immune cells, and other regulatory cells. While the regenerative medicine field is still in infancy, transplantation of functional tissues derived from patient's own cells could profoundly improve the health of patients with end-organ failure. [15]

## 6 BIOINFORMATICS

Each of the steps in analyzing disease models relies heavily on bioinformatics and big data analytic. Bioinformatics is the field combining computer science, biology, mathematics, medicine, engineering, etc. [18] When Watson and Crick first identified the DNA structure, discover quickly led to the DNA coding mechanism and the interpretation of sequencing information. The interpretation and analysis of sequencing data was very amendable to computer science. We began to sequence and interpret larger datasets including entire genes, entire chromosomes, the entire human exome, the entire human genome, and now the entire transcriptome and metabolome. Further we need to compare these large datasets to one another. Bioinformatics has gone far beyond sequence analysis to involve image analysis, mass spectroscopy, and countless other integration between biology and computer science. there are also distinct field of Biomedical informatics, which refers more specifically to the integration of computer science and medicine. This often involves running multiple subsequent computer programs in established pipelines. Projects like the Galaxy project work to streamline these pipelines for ease of use. We will discuss some common applications of bioinformatics.

### 6.1 Sequence Assembly

Sequencing technologies produce millions of fragments of DNA. Sequence assembly is the process of identifying overlapping sequence, aligning the overlapping portion and combining into a complete genome. Once the genome is assembled it is possible to compare a sample of DNA to a known sequence in a database. One of the most popular tools involves the program Basic Local Alignment Search Tool(BLAST.) Scientists can input any obtained sequence and check for matching to a known sequence in the database.

### 6.2 Sequence Annotation

Sequence annotation involves identifying the important regions in a sequence. It includes identifying the regions that code for proteins, regulatory regions, and other biologically significance sequence. It is performed by popular programs such as

### 6.3 Comparison of two states

Another set of software tools involves the comparison of two datasets. This includes the comparison of two disease states, two individuals, or any other two datasets that need comparison and analysis.

## 6.4 Examples of a Popular Bioinformatics Pipelines

The programs utilized for RNA Sequencing analysis include the Tuxedo Suite open source software package which includes Tophat, Bowtie, Cufflink, CuffCompare and CuffDiff [23]. The compressed BAM file type is utilized by these programs. Tophat aligns sequencing reads to the human genome using the high output short read aligner Bowtie and then analyzes the results to identify splice junctions. Cufflinks assembles transcripts, mapping segments of transcripts to genes and individual transcripts of a reference genome. Cufflinks uses fragment counts as a measure of relative abundance, which are reported as Fragments Per Kilobase of exon per Million fragments mapped (FPKM). Assembled transcripts from can be compared using Cuffcompare. CuffDiff to compare transcript expression level, splicing and promoter use. Cuffdiff uses the Cufflinks to compare transcript expression levels in two data sets. It allows the user to find differentially expressed and regulated genes at the transcriptional and post-transcriptional level by reporting the log-fold-change in expression.

## 7 COST OF HEALTHCARE

### 7.1 The Current State

One of the most troubling issues facing the United States, and the world, is the increasing cost of healthcare. The problems are different around the globe. Much of the developing world lacks access to adequate healthcare, which is a serious problem. This paper focuses on a different problem, in the crisis facing the United States. Current healthcare spending is greater than 3 trillion dollars [5]. This makes up 17 percent of GDP. This number grows every year and is unsustainable. This number affects citizens deeply, and currently healthcare costs are responsible for 50% of bankruptcy claims in the United States [6]. All of this extra spending does not equal better health. In most measures of health, from infant mortality to life expectancy, the United States find itself far from the top. There are major issues at play ranging from a massive bureaucracy, to the poor health and obesity of participants.

### 7.2 The Future

It is projected that the average family will spend over 25% of income on to healthcare [6]. The problem is not projected to improve. As the *baby-boomers* age, the population over 60 with high cost chronic healthcare problems, increases exponentially. In Medical School, we were taught about this *silver tsunami* approaching the US healthcare system (prompting me to go into Pediatrics.) Many individuals, including myself, look to Big Data to uncover these problems and help fix them. Before it is too late. There are technology solutions including the electronic health record, medical reference technology, genomic medicine, telemedicine, wearable health technology, and personalized medicine.

## 8 ELECTRONIC HEALTH RECORD

### 8.1 Electronic Medical Record and Genomics (eMerge)

There is currently a massive effort undertaken by multiple companies and branches of government to combine genomics data and the

electronic health record. According to the website: "eMERGE is a national network organized and funded by the National Human Genome Research Institute (NHGRI) that combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine." This method of combining genomics data and electronic health information holds great potential.

### 8.2 Adoption of and EMR

Throughout history, medical records were taken on paper, but after 2000 the slow transition to electronic records began [12]. The handwritten records were kept in large file cabinets, and when records needed to be shared between physicians or institutions (across the country or across the street), the paper records were faxed over a telephone line. This technology is decades old. As technology raced forward with supercomputers and the worldwide web, medicine continued to use these antiquated forms of communication. Finally, government mandating forced healthcare systems into the modern era and electronic records went online. Currently over 84% of health records are online [6].

### 8.3 The Current State

A majority of healthcare systems around the world are under a government regulated socialized medical system which comes with a universal health record. The healthcare system in the United States is privatized, therefore the transition to EHR came with individual health entities purchasing a multitude of different EHRs. The problem comes in that a patient presenting to two different healthcare facilities, even if across the street or within the same building, will have two different medical charts that do not communicate with one another. The other problem comes with accessing this information. The two largest companies Epic and Cerner have a commercial interest, with a primary goal to increase revenue to the shareholder. It is exceedingly difficult for the nonprofit entities including academic centers and hospitals to access the patient information within the EHR. There is tremendous potential within the EHR. Beyond data collection, storage, data retrieval, and analysis, we should move towards real time guidance and guidelines for medical decision making to improve health.

### 8.4 Phenome-wide association studies ]

The first established linkage of the electronic health record and genomics datasets took place at Vanderbilt University. Vanderbilt Medical Center began to collect biospecimens from patients (using an ethically controversial opt-out consent process.) They performed Whole Exome Sequencing on the specimens. They then linked the specimens to the electronic health record and compiled the data in a database called BioVUE. Phenome-wide association studies is the name of a method used to measure the number of phenotypes or diseases reported in the electronic health record, in relation to single nucleotide changes in the human genome [4]. Researchers can assess whether each variant is related to any disease state. The database started in 2012 and is growing rapidly. As the dataset grows, so will its power to predict disease based on single nucleotide variants. An early version of the catalog is currently available online to all individuals.

## 9 KNOWLEDGE

### 9.1 Online Genomic Resources

Most of the Genomics data is available to the public online. The National Center for Biotechnology Information (NCBI) provide a massive cache of information. Most people know about NCBI's PubMed database of over 27 million citations from biomedical literature. NCBI also hold a massive nucleotide database, with nucleotide information compiled from almost every genomic study performed to date. Their genome site holds the sequences, maps, chromosomes, assemblies, and annotations of every version of the human genome, along with mouse, drosophila, rat, EColi, Yeast, and countless other model organisms. Not only does NCBI provide a genome browser, but numerous other organizations provide this information, including Ensembl, UCSC, etc. Researchers spend hours pouring over the genome browser of their choosing, to design experiments, interpret results, and hypothesize.

### 9.2 Online Medical Resources

Only 10-20 years ago, Hospital libraries and medical school libraries were once filled with books and journal articles. If a healthcare practitioner wanted information relevant to clinical care, they went to libraries to pour through the resources with exhaustive efforts. Today, those libraries are mostly void of books. Almost every individual in Whole Exome Sequencingtern medicine has access to a computer, and usually to a handheld device, capable of accessing far more information than could ever be stored in a library. There are massive information sources, such as PubMed, and Up To Date, a point of care medical reference similar to Wikipedia, commonly used on a handheld device, with evidence based clinical guidelines contributed by over 5,000 physicians [29]. The massive amount of data now accessible to most healthcare providers and scientists is changing healthcare rapidly. Still, there is much room for improvement as care is commonly delivered based on anecdotal evidence, and cost and quality should continue to improve. Combining this online genomic information, and online medical information will provide a valuable tool to improve health.

## 10 WEARABLE TECHNOLOGY, NUTRITION AND WELLNESS APPS

Massive data sets exist, collected by insurance companies, in electronic health records, by pharmaceutical companies and genomics data sets collected by research institutions. There is another very exciting source of big data on the horizon, in personal wearable technologies, and also fitness, wellness and nutrition apps [6]. Individuals wearing FitBits, with fitness apps on their mobile devices, wearing smartwatches, etc. can track health and wellness measures in ways that once required inpatient hospital monitoring and sophisticated research lab settings. They track sleep and activity throughout the day and night. In addition, there are countless apps which track nutrition and health. People log meals and nutrition to keep accountable. Often these apps work with time tested and well researched diets including weight watchers, etc. This technology has already changed the way many individuals look at health and wellness. This exciting new dataset has great potential to advance human health and improve disease that may be the root cause of

our healthcare epidemic. Combining the massive datasets produced through wearable technology, with genomics data, holds immense potential. Measuring exercise response and sleep endurance, to nutrition and weight gain, in light of genetic background, provide incredible insight into health and disease.

### 10.1 Visual Technology

Currently procedural technology is one of the greatest expenses to the health care system. Genomics analysis holds great potential to help reduce these costs. Telemedicine involves a virtual visit between a physician and patient [9]. There are obvious benefits, especially when a patient population is spread across a wide geographic space either due to a high level of physician specialization, or a rural patient population. Highly specialized, but critical subspecialists are often in great shortage. This places a great burden on the available providers, with often unsustainable schedules. Video technology allows doctors, nurses and practitioners to visualize patients, perform a limited physical, and to communicate with individuals at a distance. There is great potential to improve cost and reduce burden. There are limitations. Many physician specialists are valued for their technical, hands on skills. Telemedicine is not much of a help, the technical procedures, such as inserting airways into the trachea of small babies, and insert central arterial lines into major vessels to deliver lifesaving medications, require hands on skills. The same goes for surgeons and other highly skilled technical professions. Interventional techniques and robotics are increasingly being used to perform procedures, but while these operations are performed, a surgeon needs to be very close, in case unforeseen accidents problems necessitate a conventional correction. Procedural specialties are the greatest expense to our healthcare system and their procedural skills are a long way from being performed through telemedicine or robotics. Genomics data will help to triage individuals, indicating response to particular treatment or technology.

## 11 COMMERCIAL GENOMICS

The company 23 and me offers genetic testing directly to consumers [31]. For around 100\$ an individual can obtain *Ancestry Services* or *Health and Ancestry Services*. Given the massive expense and resources required to analyze genomic data, the service likely provides little to no valuable information. However the market for these novelty services has exploded in recent years, as consumers grasp to understand their own genetic information. Much of the advertising, distribution, and sharing of this genetic information is done through social media. There is a multitude of health information shared over social media networks. Blogs, columns, and posts providing information about nutrition and wellness, news stories, and information sharing. The story reporting googleflis flu prediction trends ahead of the CDC, based on search history, spread virally over facebook [7]. The field will continue to expand. Soon, as technology improves, consumers will have access to their own genomics data sets. how they access in share this information is unknown.

## 12 PERSONALIZED MEDICINE

Wikipedia summarized personalized medicine as: "a medical procedure that separates patients into different groupsfi!with medical

decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease.” [28] In a way the culmination of big data and health is with personalized medicine. In a hopefully not so distant future the electronic health record, pharmaceutical data and genomic data will provide a more tailored, affordable, and high-quality approach to healthcare. The revolutions in cellular reprogramming, genome sequencing and genome editing have opened up tremendous opportunities for the study of human disease. Based on the dizzying rates of advances in the revolutionary technologies, it is not unreasonable to believe that patient-derived and genome-edited induced pluripotent stem cells models may become a dominant model for the study of disease and the search for new therapies.

Whole Exome Sequencing and Whole Genome Sequencing can be utilized to measure all genomic changes, and newer technologies allow us to perform personalized omics measurements in affected tissue including metabolomics, transcriptomics, proteomics, etc. For example, we can take a patient blood sample, derive cardiomyocytes, neurons, smooth muscle, etc, and perform analysis to measure tissue metabolics, RNA transcriptional differences, and pharmacologic response of the tissue. At some point in the future we may move toward autologous transplantation with genetically edited organs derived from the patients own tissue. Bioinformatics analysis and interpretive steps lag behind. Clinically actionable results would be needed in hours to days, versus the months this type of analysis usually require. This rapid analysis is a rate limiting step, but is improving exponentially.

### 13 CONCLUSION

As the population continues to grow, we will continue to utilize and increasing amount of resources. Optimal utilization of these resources is the only way to ensure survival and proper living standard for the human population. Many look to the revolutions in genomic medicine combining this omics data with the electronic health record, wearable technology, pharmaceuticals and procedures to move us towards personalized, precision, medicine. Big Data is plays an increasing role in sustaining and improving our world.

### ACKNOWLEDGMENTS

Thank you to Dr. Geoffrey Fox, Gregor von Laszewski, and all of the course instructors for an excellent introduction to Big Data and Data Science.

### REFERENCES

- [1] [n. d.]. ([n. d.]). Vanderbilt University: Introduction to Bioinformatics Course Lectures.
- [2] Rudolf Amann and Bernhard M Fuchs. 2008. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology* 6, 5 (2008), 339–348.
- [3] Francis S Collins, Michael Morgan, and Aristides Patrinos. 2003. The Human Genome Project: lessons from large-scale biology. *Science* 300, 5617 (2003), 286–290.
- [4] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. 2010. PheWAS: demonstrating the feasibility of a phenotype-wide scan to discover gene-disease associations. *Bioinformatics* 26, 9 (2010), 1205–1210.
- [5] Centers for Medicare & Medicaid Services et al. 2014. National health expenditures 2012 highlights. *Online verfügbar unter* <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/National-HealthExpendData/Downloads/highlights.pdf> (2014).
- [6] Geoffrey Fox. [n. d.]. Unit 6 Lectures. ([n. d.]).
- [7] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.
- [8] Angelika Görg, Walter Weiss, and Michael J Dunn. 2004. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* 4, 12 (2004), 3665–3685.
- [9] Maria Hernandez, Nayla Hojman, Candace Sadorra, Madan Dharmar, Thomas S Nesbitt, Rebecca Litman, and James P Marcin. 2016. Pediatric critical care telemedicine program: A single institution review. *Telemedicine and e-Health* 22, 1 (2016), 51–55.
- [10] Dirk Hockemeyer, Frank Soldner, Caroline Beard, Qing Gao, Maisam Mitalipova, Russell C DeKelver, George E Katibah, Ranier Amora, Elizabeth A Boydston, Bryan Zeitler, et al. 2009. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology* 27, 9 (2009), 851–857.
- [11] Robin Holliday. 2006. Epigenetics: a historical overview. *Epigenetics* 1, 2 (2006), 76–80.
- [12] Erik WJ Kokkonen, Scott A Davis, Hsien-Chang Lin, Tushar S Dabade, Steven R Feldman, and Alan B Fleischer. 2013. Use of electronic medical records differs by specialty and office settings. *Journal of the American Medical Informatics Association* 20, e1 (2013), e33–e38.
- [13] Pauline C Ng and Ewen F Kirkness. 2010. Whole genome sequencing. In *Genetic variation*. Springer, 215–226.
- [14] Emily Niemitz. 2007. Variants of unknown significance. *Nature Genetics* 39, 11 (2007), 1313–1314.
- [15] Adrian Ranga, Nikolche Gjorevski, and Matthias P Lutolf. 2014. Drug discovery through stem cell-based organoid models. *Advanced drug delivery reviews* 69 (2014), 19–28.
- [16] Jorge S Reis-Filho. 2009. Next-generation sequencing. *Breast Cancer Research* 11, 3 (2009), S12.
- [17] HJ Rippon and AE Bishop. 2004. Embryonic stem cells. *Cell proliferation* 37, 1 (2004), 23–34.
- [18] Iwan Saeyns, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 19 (2007), 2507–2517.
- [19] Carol Jean Saunders, Neil Andrew Miller, Sarah Elizabeth Soden, Darrell Lee Dinwiddie, Aaron Noll, Noor Abu Alnadi, Nevene Andraws, Melanie LeAnn Patterson, Lisa Ann Krivohlavek, Joel Fellis, et al. 2012. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science translational medicine* 4, 154 (2012), 154ra135–154ra135.
- [20] Stephan C Schuster. 2008. Next-generation sequencing transforms today's biology. *Nature methods* 5, 1 (2008), 16–18.
- [21] Cory Smith, Athurva Gore, Wei Yan, Leire Abalde-Artistain, Zhe Li, Chaoxia He, Ying Wang, Robert A Brodsky, Kun Zhang, Linzhao Cheng, et al. 2014. Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. *Cell stem cell* 15, 1 (2014), 12–13.
- [22] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *cell* 131, 5 (2007), 861–872.
- [23] Cole Trapnell, Lior Pachter, and Steven L Salzberg. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 9 (2009), 1105–1111.
- [24] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. 2012. Five years of GWAS discovery. *The American Journal of Human Genetics* 90, 1 (2012), 7–24.
- [25] Zhong Wang, Mark Gerstein, and Michael Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* 10, 1 (2009), 57–63.
- [26] Ronald J Wapner, Christa Lese Martin, Brynn Levy, Blake C Ballif, Christine M Eng, Julia M Zachary, Melissa Savage, Lawrence D Platt, Daniel Saltzman, William A Grobman, et al. 2012. Chromosomal microarray versus karyotyping for prenatal diagnosis. *New England Journal of Medicine* 367, 23 (2012), 2175–2184.
- [27] James D Watson, Francis HC Crick, et al. 1953. Molecular structure of nucleic acids. *Nature* 171, 4356 (1953), 737–738.
- [28] Wikipedia. [n. d.]. Personalized Medicine. ([n. d.]). [https://en.wikipedia.org/wiki/Personalized\\_medicine](https://en.wikipedia.org/wiki/Personalized_medicine)
- [29] Wikipedia. [n. d.]. UpToDate. ([n. d.]). <https://en.wikipedia.org/wiki/UpToDate> Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 2 Sept. 2016.
- [30] Yaping Yang, Donna M Muzny, Jeffrey G Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, et al. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine* 369, 16 (2013), 1502–1511.
- [31] Patricia J Zettler, Jacob S Sherkow, and Henry T Greely. 2014. 23andMe, the Food and Drug Administration, and the future of genetic testing. *JAMA internal medicine* 174, 4 (2014), 493–494.

- [32] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics* 10 (2009), 451–481.

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no key, author in vanderbilt
Warning--no author, editor, organization, or key in vanderbilt
Warning--to sort, need author or key in vanderbilt
Warning--no key, author in vanderbilt
Warning--no key, author in vanderbilt
Warning--no key, author in vanderbilt
Warning--no author, editor, organization, or key in vanderbilt
Warning--empty author in vanderbilt
Warning--empty year in vanderbilt
Warning--no number and no volume in centers2014national
Warning--page numbers missing in both pages and numpages fields in centers2014national
Warning--empty year in fox6
Warning--empty address in ng2010whole
Warning--empty year in wiki-personalized
Warning--empty year in wiki-updated
(There were 15 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.36.22] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
```

```
Typesetting of "report.tex" completed in 1.1s.  
..../README.yml  
24:14      warning  truthy value is not quoted  (truthy)  
28:25      error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Durbin, Matthew  
hid: 311  
paper1: 100%  
paper2: in progress  
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)  
wc 311 project (null) 6591 report.tex  
wc 311 project (null) 7166 report.pdf  
wc 311 project (null) 1117 report.bib
```

```
find "
```

---

```
105: In 1977 a paper was published entitled "A rapid method for  
determining sequences in DNA by primed synthesis with DNA  
polymerase". This technique, now known as Sanger sequencing,  
revolutionized molecular biology. Using termination of sequence  
and dye detection, it provided a fast and easy way to determine  
the DNA sequence of living organisms. It is still extensively  
utilized. Many newer technologies have been developed and are  
known as "next generation sequencing." \cite{schuster2008next}
```

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0
```

```
tables 0
```

```
includegraphics 0
```

```
labels 0
```

```
refs 0
```

```
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no key, author in vanderbilt
Warning--no author, editor, organization, or key in vanderbilt
Warning--to sort, need author or key in vanderbilt
Warning--no key, author in vanderbilt
Warning--no key, author in vanderbilt
Warning--no key, author in vanderbilt
Warning--no author, editor, organization, or key in vanderbilt
Warning--empty author in vanderbilt
Warning--empty year in vanderbilt
Warning--no number and no volume in centers2014national
Warning--page numbers missing in both pages and numpages fields in centers2014national
Warning--empty year in fox6
Warning--empty address in ng2010whole
Warning--empty year in wiki-personalized
Warning--empty year in wiki-updated
(There were 15 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
non ascii found 8211  
non ascii found 8217  
non ascii found 8212
```

```
=====  
The following tests are optional  
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

```
-----  
passed: True  
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

```
-----  
passed: True
```

# **Big Data Mental Health Monitoring: A Private and Independent Approach**

Neil Eliason  
Indiana University  
Anderson, Indiana 46017  
nreliaso@iu.edu

## **ABSTRACT**

Big data holds great promise in a number of fields, and mental health treatment is no exception. Effective big data applications have been developed for every stage of treatment, but not without difficulties. This is particularly true for passive monitoring using smartphones. While it provides improved resolution of consumer behavior, it has complications with privacy and implementation connected with use of commercial software and streaming data. This project created an open source program which uses discrete exported data from smartphones to implement passive monitoring in a way that promotes consumer control of their data and independence from commercial interests. A working program which parses and analyzes smartphone data was created to demonstrate concepts, but would benefit from further development of parser options, analyses, and input/output file management in the future.

## **KEYWORDS**

i523, passive monitoring, open source, mental health treatment, big data

## **1 INTRODUCTION**

Can big data contribute to the treatment of mental illness? Are there any unique issues that hinder big data being used for mental health treatment? Big data has certainly gained the reputation of being a formidable remedy for analytical problems. Corporate empires such as Google, Facebook, and Amazon were built on a foundation of handling and managing big data, and scientists have explored everything from the structure of galaxies to the language of our genes using the power of big data. Perhaps it can work such wonders in the field of mental health as well.

### **1.1 What is Big Data?**

First of all, what is big data? This proves to be a difficult question to answer, as definitions vary based upon industry and often shift as technology evolves and adapts. However, there are three generally accepted traits of big data: high volume (amount of data), high velocity (rate of data creation), and/or high variety (number of data sources). It is when these data traits become increasingly extreme, and require non-traditional analytic methods to extract insights, that it becomes big data [9].

The need for big data analytics has arisen, partly because data storage capabilities have increased at a faster rate than that of data processing. This paired with the proliferation of data collecting devices creates a surplus of stored data that is growing faster than it can be processed traditionally. This drives the demand to develop new ways to extract insights from big data [6].

Big data approaches can be applied to every stage of the data life cycle. Unworked data can be extracted from varied and possibly streaming sources and cleaned and organized automatically. Then predictive analytics can be applied to identify patterns in the data, relying on either regression techniques or machine learning. These approaches attempt to scale these extreme data sets to the level of human insight by reducing noise and identifying patterns via automation and intelligent algorithms [9].

### **1.2 Mental Health Treatment**

Mental illness has been a prevalent issues for all societies worldwide. It has been estimated that 29.2 % of the world population will personally struggle with mental illness, and that 17.6 % had a mental illness in 2014 [30]. In the United States, 17.9 % of the population was estimated to have mental illness [21]. The negative impact of these disorders is high. It is estimated by the Center for Disease Control and Prevention that 36,035 Americans died by suicide and 666,000 visited acute care for harm to self in 2008 [7]. 1,947,775 Americans drew social security/disability due to a psychotic or mood disorder according to the Social Security Administration in 2013, making up around 19 % of recipients [29]. In 2002, mental health issues are estimated to have had \$ 100 billion negative effect on the US economy [21], and over 12,000 facilities were providing mental health services in 2015 in the United States [31]. It is evident given the scale of the negative impact mental illness has on individuals and societies, that effective solutions are needed.

Services provided to meet this need vary depending on a variety of factors, but typically involve a screening and assessment process [1], assignment of interventions [32], and monitoring of treatment progress [10]. Mental health screening is an initial contact that takes a relatively succinct amount of information and seeks to guide the person towards the right services. They are designed to not be time consuming or overly invasive to distribute them to a wide group of people. Assessment is more comprehensive, with the goal of identifying primary mental health needs and clinical diagnoses for the purpose of informing treatment decisions [1].

After the treatment team has determined a person's needs and diagnosed clinical disorders, appropriate interventions are chosen, and added to the person's treatment plan. Typical services are talk therapy which targets developing positive change strategies, medication management which seek to manage symptoms through psychiatric drugs, and case management and related support services which assist in coordinating details of their care and applying skills learned in treatment [32].

Treatment monitoring is where clinicians assess if treatment is resulting in positive change for the person receiving treatment.

Without this feedback, it is easy for clinicians to lose sight of how the person is doing. Though most clinicians have methods of assessing progress, it can be difficult to do so objectively [10].

The majority of the activities of mental health treatment involve gathering data about the treatment consumer and extracting clinically meaningful insights. This frequently generates considerable amounts of data, which is often from a variety of sources, thus making big data approaches appropriate for consideration.

### 1.3 Big Data Mental Health Applications

Given the problem-solving potential of big data analytics, many researchers have explored ways to apply these techniques to the problem of treating mental health difficulties. Providing quality mental health treatment involves considerable information gathering and insight extracting work, thus making big data techniques relevant for every stage of the treatment process.

*1.3.1 Screening and Assessment.* Mental health screening typically is the first contact people have with mental health services, and serve the important role of identifying mental health needs at a larger scale. Thus screening methods that are easy to implement and capture information from a large group are ideal. Many big data approaches to this problem have been attempted, often using data found on social media. This provides a large, if not messy dataset, but the accuracy of these methods was found to be better than that typical of primary care providers, but worse than self-reporting tools [11].

Assessment and diagnosis is more information intensive and traditionally requires the skills of a highly trained clinician. Many studies have looked at how to streamline that task by using data gathered through data mining and natural language processing to group people into diagnostic categories using machine learning techniques. While this approach has some success, it has not demonstrated more accuracy than traditional assessment methods, thus suggesting that it may primarily serve an assisting role for the time being. Related to big data assessment techniques, outcome prediction using machine learning has been used to attempt to identify a person's likely treatment trajectory, given certain factors. These predictive models sought to connect risk factors with negative outcomes, and was able to do so with a fair degree of accuracy (69% to 99%). However, sample sizes were small, so further research is needed to fully assess their efficacy [15, 19].

*1.3.2 Interventions.* Treatment interventions are the actual delivery of services, such as therapy or medication management. They are very personalized and focus on facilitating positive change in the treatment consumer. Thus there are not as many Big Data opportunities, as intervention itself does not generate massive amounts of data. However, certain web-based interventions could use big data techniques by delivering interactive services to a large number of consumers at one time, thus requiring specialized analytic techniques to respond to user input [17, 18].

*1.3.3 Treatment Monitoring.* Just as screening and assessment provide the clinician with information to guide what treatment they are to receive, treatment monitoring aims to inform the clinician on the efficacy of treatment. Though this is just as important, it is often difficult to be objective and to engage the consumer in

providing needed information. Active monitoring via smartphones has been explored as a possible solution. Through apps or text messages, a consumer is reminded of treatment goals and symptoms are assessed in real time. This data can be generated multiple times a day, and takes a variety of forms, which indicates a possible big data opportunity [18].

Passive monitoring is a similar approach, but instead of relying on consumers to actively respond, it gathers data from them throughout the day. This data can be collected in a number of ways, including smartphones and wearable devices. Many studies have paired this active monitoring data with machine learning techniques to create predictive models [18]. One example of this is an initiative to develop a program which can identify whether someone is experiencing symptoms of bipolar disorder from input on their smartphone such as data from the devices sensors or from keyboard input [34]. The app is being developed in Apple's ResearchKit, which is an open source medical research application development tool [3].

While these techniques are promising, implementing them can be challenging, given the variety of data sources involved from different devices. It is also difficult to test the approaches with consumers, due to lack of engagement [19]. Provision of technical support and clinical engagement concerning passive monitoring techniques has shown to help improve consumers' level of engagement [28].

### 1.4 Barriers to Big Data for Mental Health Treatment

Every stage of treatment can benefit from big data applications to differing degrees. The screening and assessment process is the most information intensive stage, and thus has received considerable attention from Big Data application research. These efforts have had some success, though they have not surpassed traditional methods, and have not been validated on larger samples sizes [19].

However, treatment monitoring, particularly passive monitoring is considerably information intensive as well, and can potentially produce datasets with more volume, variety, and velocity than initial assessment services. As seen above, development of these approaches appears to be slower, due to a number of issues inherent in many Big Data applications, which are accentuated in passive monitoring. For this reason development of this approach has been slower, though it has elicited considerable interest and discussion [18].

A primary concern is that privacy will be compromised for persons who participate in treatment that utilizes big data techniques, particularly passive monitoring. The use of commercial software for data analysis often requires that the analytics company process the data themselves rather than the treatment provider. This is especially true of mobile device apps, which usually take data, and send it back to their own servers to be processed. These applications do not necessarily have the same privacy rules that medical records due concerning protected health information [19]. A key part of privacy is one's ability to control what information about them goes where, and to prevent unwanted information from being shared [14]. With streaming personal data from smartphones, consumers begin lose some of their privacy, because they lose control of their

data. As it is constantly being sent to the treatment provider, and requires action on the part of the consumer for it to stop, the person has sacrificed some of their privacy in order to receive this service. Though this is a common trade off in medical and mental health contexts, research indicates that people prefer to have more control of the private data, and that they want to be able to share it in portions, rather than have to share all of it [4].

Another issue related to using activity monitoring in mental health treatment is that the variety of competing smartphone and wearable sensor companies creates an environment with plenty of human behavior Big Data, but it is not easy to integrate. This data is stored on separate private databases and each company has fiscal motivations to resist collaboration. Thus the product which a treatment provider intends for use as a clinical tool, is also being used for a commercial purpose, and to create a comprehensive application which is not encumbered by the independent economic interest of a private business is difficult [12].

Some companies attempt to avoid this by providing some open source products. Open source software is freely distributed, can be modified and integrated into other software, the source code is available, and it is not associated with any specific product. Benefits of such code is that they can be improved by a large number of programmers, they can be widely implemented, and it is not tied to any product or corporation [22]. An example of such as Apple's ResearchKit [3]. However, though the product is free, it is still tied to the company's resources, in this case Xcode [33]. Though some development can be done freely, Apple controls this, and extensive work cannot be easily distributed without paying for a developer account [2]. While there are some truly free open-source software solutions for mental health, they tend to target administrative problems, rather than treatment itself [16].

## 1.5 Open Source and Discrete Data Transfer

A true open source big data solution could make passive monitoring an ethical and workable tool for mental health treatment. A program written in the open source programming language Python would be free of corporate entanglements and costs, but would benefit from the massive amount of documentation, packages, and ready-made code produced by the Python community [26]. Run on open source Ubuntu Linux [5] installed on open source Oracle Virtual Box [23], such a program would be independent of commercial interests, completely reproducible, and free of charge. Besides these direct benefits, open source programs often have performance and security advantages of commercially developed products [20].

The data source for the passive monitoring would still be smartphones, but instead of using commercially provided apps for analysis, the data would be extracted and analyzed using the open source based program. The iPhone step count data collected by the built-in accelerometer, and stored in the Health App can be exported via email as an xml file [24]. Though this method losses the benefits of streaming data, it increases the consumers' control over their data by making data transfer definite and discrete. Instead of agreeing to install an app on their device which will track their behaviors forever unless they ask for data to stop streaming or the app is uninstalled, the consumer can agreed to provide a defined amount of information now, and may do so again later. As the Health export

file stores all detected steps, no data is lost, it is just not seen in real-time. This would still provide useful information for treatment monitoring purposes.

Using this method also allows for the steps data of numerous people to be parsed and analyzed in an automated fashion, which would be necessary from a mental health treatment perspective, as numerous consumers data would need processed. As datasets increase, some way of address the increasing extremity of the data needs to occur. There are other approaches that would be insightful, but this approach is a good fit when the output is a file for each individual participant, rather than aggregate date about a group.

## 1.6 Thesis

Private and independent passive monitoring can be utilized in mental health treatment by leveraging open source programming tools to analyze aggregate movement data provided from smartphones in discrete amounts. This approach avoids the cost and entanglements of commercial software and wearable technology, as well as increases consumer control over their personal data.

## 1.7 Project Goals

This project attempts to demonstrate this concept by developing a simple open-source program written in Python which can perform full data life cycle analytics on automatically collected iPhone Health App data. The program was designed to accommodate Big Data by automatically iterating over multiple files without user input.

## 2 METHODS

### 2.1 Design

This current research utilized the Python programming language to develop a program which could parse, analyze, and report clinically relevant information from a folder of exported individual iPhone 6 Health App data files. This program consisted of four sub-programs: acceleparser.py, Tables.py, Visualizations.py, and makefile.py. Also included with the program code was a folder named iPhoneData which contained the test xml files, and a bash script named make\_install.sh which installed program packages.

The acceleparser.py program which imports the xml file and parses it using the ElementTree python package. The xml file is imported and parsed into a tree structure. It then iterates through the tree and appends date/time data and steps data to corresponding lists. Those lists are then used to create a dataframe using the pandas python package, which is then returned.

The Tables.py program takes the dataframe returned from acceleparser.py, and formats it for display and for use by the Visualizations.py program. Tables.py consists of two functions, stepsBYdata and stepsBYweekday. The stepsBYdata function formats the dataframe to show the total steps for each date using the pandas groupby functionality. The stepsBYweekday function formats the dataframe to display the mean steps for each date in columns of weekdays, and rows of weeks labelled by the first Monday's date. This was accomplished using pandas pivot table functionality to aggregate the mean function over the dataframe, and then a new dataframe was made with the weekday columns. Each function returns a dataframe object.

The Visualizations.py program takes input dataframes from either acceleparser.py or Tables.py, and returns graphs. Visualizations.py consists of two functions. The stepsBYdateGraph function takes the dataframe created by the stepsBYdate function of Tables.py and creates a time series line graph using matplotlib.pyplot python package. The stepsBYweekdayGraph function takes the dataframe returned by acceleparser.py, and creates a list of mean steps for each day of the week and a list of days of the week. From these lists, a bar graph of the mean of steps for each day of the week is generated using matplotlib.pyplot.

The makefile.py program iterates over the xml files stored in the iPhoneData folder, and for each file ran the acceleparser.py program, and directed that dataframe to the Tables.py and Visualizations.py programs to output the table and graphs. The table was then saved as a txt file and the graphs saved as pdf, each named by which iteration the original file was in the for loop.

This program utilized the module design of small sub-programs to provide ease of customization and expansion of program features. Later functions can be added to their appropriate subprograms, and the makefile.py modified to include the new function, and thus generate a new report. Multiple reports could even be added to the makefile.py, and executed when called.

## 2.2 Data

The test data consisted of xml files exported from the Apple Health App on iPhone 6. To export the data, the export health data option was selected in the app, which compiled as export.zip. This file was then emailed to researcher and the file unzipped as folder named apple\_health\_export. This folder contained two files, export.xml and export\_cda.xml. The export.xml file contained the steps data, and was renamed “Client1.xml”. It was then placed in the folder “iPhoneData” in the github repository [8].

Health data was exported from two devices, which contained a variety of data, but only date/time and step count data were used in this project. The xml files were 5.58 MB and 8.33 MB in size, containing 203 days and 888 days of steps data respectively. Data was also tested from iPhone 8 models, but the program would not parse the files, and thus were excluded from the final test set.

## 2.3 Analyses

Basic descriptive statistics were utilized to explore patterns of step activity over time. Daily step counts were organized in table fashion by day of the week columns and rows of weeks. Visual analyses consisted of a bar graph of the mean steps taken for each day of the week and of a time series line graph of daily steps taken over time. These analyses were chosen to show basic patterns in data in a way which could be quickly assimilated.

## 2.4 Specifications

Project development and testing was done in Ubuntu Linux operating system version 16.04 (download available at [5]) installed on an Oracle Virtual Box Graphical User Interface (download available at [23]). Code was written in Python version 3.5.2 (download available at [27]) installed within pyenv Python Version Management System (download available at [36] per installation instructions available here [35]). External Python libraries utilized in project were pandas,

matplotlib, and numpy (download available from the Python Package Index [25]). With this configuration, it was necessary to install tk-dev system wide prior to creating pyenv virtual python environment (instructions found at [13]), and to install python3-tk within the virtual python environment in order to utilize matplotlib.pyplot. Completed source code can be found at Neil Eliason’s bigdata-i523 github repository [8].

## 2.5 Procedure

Program tests were done per instructions found in the README.md file of the project sourcecode [8].

## 3 RESULTS

After executing the program, the output was one txt file and four pdf files for every one input xml file. Thus, for the test data of two xml files, ten files were created. All files were generated in the code directory from which makefile.py was ran.

The txt file contained a table with columns of daily steps for each day of the week and rows labelled as the date of the first day of the week, with each week starting on Monday. It also contains the spearman correlation of daily steps with day of the week. The file name was determined by where in the order the original xml data was parsed in the program iteration. For example, the output txt file for the first parsed xml file would be named “Client1.txt”.

**Table 1: Output table of daily steps by day of the week**

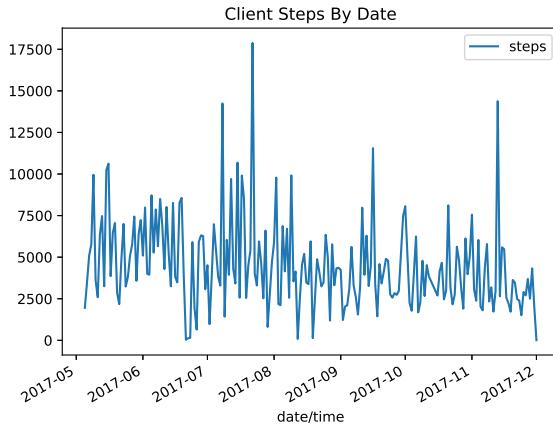
Client1 Report							
Week Of	Steps by Week						
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
2017-05-08	[5766]	[9956]	[3605]	[2590]	[1955]	[3471]	[5097]
2017-05-15	[10236]	[10624]	[3861]	[6464]	[6358]	[7473]	[32511]
2017-05-22	[4884]	[6999]	[3233]	[3796]	[7056]	[2858]	[2174]
2017-05-29	[3579]	[6357]	[7232]	[5091]	[5076]	[5761]	[7443]
2017-06-05	[8717]	[5276]	[7870]	[5658]	[7986]	[3999]	[3950]
2017-06-12	[8018]	[5195]	[3246]	[8261]	[8501]	[6930]	[4284]
2017-06-19	[8568]	[3871]	[33]	[127]	[3811]	[3484]	[8264]
2017-06-26	[645]	[5922]	[6309]	[6272]	[148]	[5905]	[1938]
2017-07-03	[3831]	[6986]	[5413]	[3823]	[3074]	[4517]	[974]
2017-07-10	[6040]	[3938]	[9701]	[4309]	[3292]	[14241]	[1423]
2017-07-17	[9911]	[8529]	[2545]	[4485]	[3416]	[10684]	[2562]
2017-07-24	[3296]	[5949]	[4651]	[2529]	[5423]	[17880]	[4017]
2017-07-31	[4718]	[5910]	[9783]	[2178]	[6599]	[802]	[4146]
2017-08-07	[6714]	[2555]	[9921]	[3544]	[2115]	[6874]	[2536]
2017-08-14	[4568]	[5198]	[3473]	[3378]	[4136]	[83]	[2455]
2017-08-21	[4877]	[4141]	[3241]	[3481]	[5952]	[134]	[1186]
2017-08-28	[5774]	[3310]	[4330]	[4362]	[6347]	[4951]	[2028]
2017-09-04	[2102]	[3124]	[5619]	[3286]	[4236]	[1219]	[3186]
2017-09-11	[7976]	[3939]	[6293]	[3262]	[2610]	[1545]	[3437]
2017-09-18	[1442]	[4586]	[3420]	[4111]	[4484]	[11554]	[2759]
2017-09-25	[2564]	[2841]	[2733]	[2959]	[4901]	[4790]	[8063]
2017-10-02	[5124]	[2248]	[1774]	[4046]	[5142]	[7477]	[2276]
2017-10-09	[4787]	[2663]	[4525]	[3840]	[6244]	[1681]	[3175]
2017-10-16	[2692]	[4149]	[4667]	[2463]	[2988]	[8118]	[6127]
2017-10-23	[2163]	[2744]	[5631]	[4805]	[3134]	[1902]	[2025]
2017-10-30	[3976]	[5048]	[7557]	[3065]	[2378]	[6034]	[2963]
2017-11-06	[1800]	[4362]	[5793]	[2332]	[3186]	[1723]	[1719]
2017-11-13	[14382]	[2648]	[5592]	[5482]	[2561]	[2230]	[2704]
2017-11-20	[3631]	[3463]	[2485]	[2390]	[1509]	[2906]	None
2017-11-27	[3691]	[2503]	[4328]	[2005]	[11]	None	None

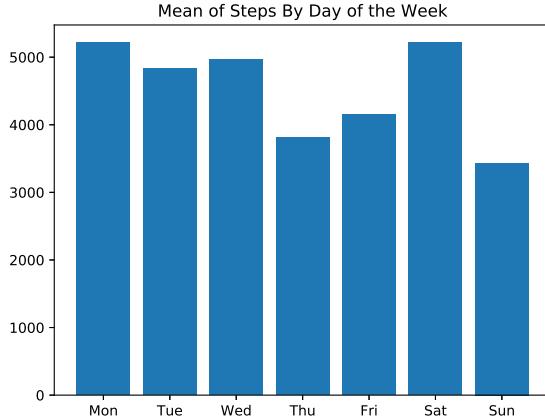
Correlation of Steps with Day of the Week	
weekday	steps
1.000000	-0.198758
steps	-0.198758 1.000000

The pdf files were graphs of different analytics and relationships of the parsed data. The first was a time series line graph of daily step information for the whole time period of the xml file. The second was a bar graph of the mean steps for each day of the week

of the whole time period of the dataset. The third was a bar graph of the standard deviation of daily steps for each day of the week. The fourth was a scatterplot with points representing number of daily steps on the y axis and the day of the week on the x axis. These files were named similarly to the above txt file, with each file named according to its order in the iteration. For example, the first xml file parsed would generate “Client1StepsByDate.pdf” for the first graph and “Client1MeanStepsByDayOfWeek.pdf” for the second, “Client1StDvStepsByDayOfWeek.pdf” for the third, and “Client1ScatterplotOfStepsByDayOfWeek.pdf” for the fourth.



**Figure 1: Output graph of time series of daily steps**

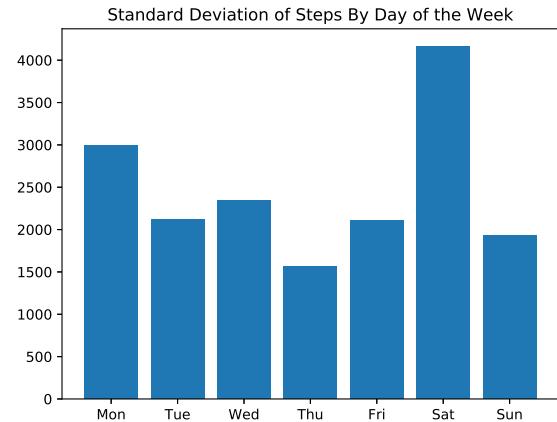


**Figure 2: Output graph of mean steps for each day of the week**

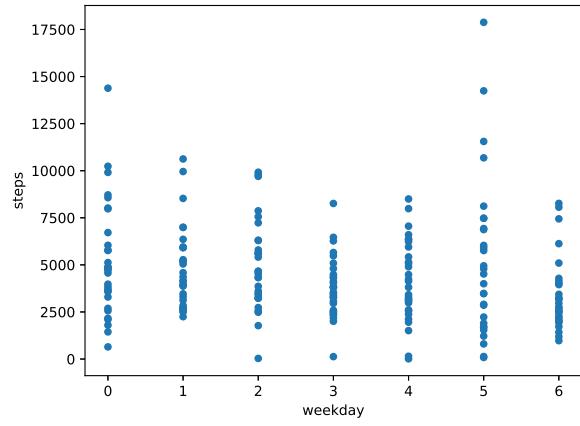
## 4 DISCUSSION

### 4.1 Project Goal Assessment

The goal of this project was to address the common issues of privacy concerns and ease of implementation that hinder applying Big Data



**Figure 3: Output graph of standard deviation of steps for each day of the week**



**Figure 4: Output graph of scatterplot of steps by each day of the week**

techniques to Mental Health Treatment. The strategy was to write a program that leveraged open source software, used smartphone data provided by the consumer in discrete and definite ways, and allowed for applications in a Big Data setting.

**4.1.1 Open Source.** The source code for the program was written in Python 3.5.2, and utilized the pandas, matplotlib, and numpy packages to perform data cleaning, organization, and visualization. All these resources are available online free of charge, and with substantial technical support and documentation. By following the installation instructions found in the README.md, this program can be installed on the open source Ubuntu operating system on Virtualbox, thus making the program configuration reproducible on any standard operating system with sufficient hardware. This allows for distribution in various settings with no cost or input

from outside commercial interests. The project successfully created a working program using open source programs.

**4.1.2 Discrete Data Delivery.** Data was exported from the iPhone Health app, and emailed to the researcher as a single file. By using a discrete dataset, the person providing the data did not have to install invasive apps on their phone or agree to releasing information for an indefinite amount of time. They made the decision to provide a set amount of data, and can make a subsequent decision to do so at a later time, but they do not have to make a decision to stop sharing the data. They must be active to share in the future, rather than needing to be active to stop sharing. The project successfully utilized discrete data as the data source for the program.

**4.1.3 Big Data Informed.** By utilizing a design that iterates the program through a folder of data files, one command from the terminal can theoretically produce the output files for hundreds of consumers. This functionality is necessary when dealing with big data, as it would be incredibly tedious, if not impossible to do such a task through the graphical user interface. While the program was designed to be useable with big data, testing was not done with a dataset which would qualify as big data. Also, no big data analytics, such as machine learning were utilized. The project successfully developed a starting platform for big data analytics, but needs further development.

## 4.2 Thesis Assessment

This project was able to create an open-source program which analyzes readily available behavior data from smartphones provided in discrete samples, rather than streaming. While the project goals were met, how do they relate to the thesis? Do these results support the thesis that open-source programs and discrete data collection address issues of consumer data control and flexible implementation of passive monitoring in mental health?

The diverse selection of technology companies fighting for a larger piece of a lucrative market share certainly creates numerous products which could be used for passive monitoring applications. However, corporate competition hinders collaboration and integration of these tools into a comprehensive mental health Big Data approach [12]. Also, by utilizing commercial software, data is not always guaranteed to be protected, and may be gathered from apps and analyzed by the corporation for business purposes [19]. By utilizing completely open source products for the project, to analyze multiple persons' accelerometer data from a popular smartphone device, this project demonstrates that a corporate interference free approach to passive monitoring is possible. By avoiding corporate interference, the implementation of this approach is completely flexible and the fate of data is clear and transparent encouraging safety of consumer data.

Much attention has been given to the power of streaming data, which is not unwarranted. In the context of passive monitoring, streaming data can provide an incredible level of information resolution about a consumer's behavior [12], and safety planning and monitoring could utilize such approaches to identify when someone may be more symptomatic and in danger [18]. However, these approaches to passive monitoring remove control of the data from the consumer, and force them to take action if they want that data

sharing to stop. Given that people generally prefer to have greater control over what data is shared and when [4], this may contribute to some of the lack of engagement found in some implementations [18]. By using discrete streaming data, this project allows for consumers to have more control over their personal data, while still benefiting from the clinical insights afforded by passive monitoring. Another ancillary benefit of utilizing discrete data transfer is that it actually facilitates more interaction between the consumer and the service provider about the health data. This can provide opportunities for insight development for the consumer by increasing awareness of their behaviors and activity.

## 4.3 Limitations and Future Directions

This project's aim was to develop a prototype passive monitoring program using open source and discretely provided consumer data that could be utilized with Big Data. While this was accomplished, the actual work the program does is fairly basic. The hope would be for more robust and effective models of passive monitoring to be built expanding on these approaches. As it is, the current project has the following limitations.

**4.3.1 Depth of analysis.** As seen above, a number of powerful analytic techniques are available for use on big data sets. From traditional inferential statistics, to machine learning, to advanced visualizations, more extensive analytics could provide increased insight from this data. This project included a simple spearman correlation to demonstrate how inferential analysis modules can be added to the functions of the program, and thus provide more in-depth insights. By including another variable, such as reported mood, hours of sleep, etc. machine learning techniques could attempt to identify patterns between the variables. This particular data may especially benefit from exploring patterns at different levels of resolution, attempting to identify patterns in activity based on time, day, week, or month, and pair this with more information rich and refined visualizations.

**4.3.2 Narrow data sources.** As the project progressed, it became apparent that the format of the xml files differed between iPhone versions. The xml parser program which was written using iPhone 6 test data, would not read the iPhone 8 data. The utility of the program would be greatly increased if it was able to read the xml files from other iPhone models. Given the module design of the program, a smart parser could be developed, which would identify which version of iPhone exported the file, and have utilize a parser algorithm which is compatible with the xml structure. Further work could even explore creating an Android parser module, which could create compatible dataframes for the subsequent analysis modules.

**4.3.3 Requires some technical knowledge.** Though to implement this program requires no ability to code, the creation of a Virtual Box, installation of Ubuntu, and executing commands via terminal would be difficult for a novice with no guidance. This could be problematic for real-world implementation, and many mental health providers may not feel comfortable using the commandline. Future implementations would benefit from detailed documentation with the technical layman in mind, in order to facilitate utilization by clinical staff who may not be familiar with the technology. Another approach would be to develop a more comprehensive installation

script and graphical user interface for those who are not comfortable with command line.

*4.3.4 Input and output files could be more organized.* The program as developed does not place the output files in any specific location, but rather allows them to generate within the code folder. While workable, this could become a bit cluttered as large numbers of files are generated. Also the program's file naming technique consists of assigning a number to the consumer's data based on which iteration of the program in which it is parsed. This means that in order for the identity of the person to remain connected to their data, care must be taken to order the files correctly in the iPhoneData folder. Otherwise, if the xml file for Client1 is accidentally put third in the iPhoneData folder, then that record and all following records will be connected to the wrong person. One possible solution is to utilize a database program, perhaps the open source SQL database MySQL, to manage the original files and their relationships to consumer identifying information and output files. This could prevent the likelihood of errors occurring, and also introduce some of the functionality brought to bear from the SQL programming language.

## 5 CONCLUSION

Big Data techniques have demonstrated great success in a variety of fields, but can they positively impact the provision of mental health treatment? The question is an important one, given the prevalence of mental health difficulties worldwide, and the large cost they have on individuals and societies.

Researchers have explored big data approaches for every stage of the treatment process, and found effective applications, such as using social media to detect depression, assigning diagnostic criteria using machine learning, or detecting a manic episode by the way a consumer is typing. The research focuses on the more information intensive areas of screening/assessment and treatment monitoring, which lend themselves to Big Data analysis. With the prevalence of smartphones and wearable devices which people take with them throughout the day, treatment monitoring is of particular interest. Active monitoring approaches require intentional interaction from the consumer and passive monitoring gathers data from the various sensors and inputs of the device without any intentional action from the consumer.

While these technologies have great potential, concerns of privacy and effective implementation hinder their growth. Many commercial methods exist for activity monitoring, but they do not integrate well with each and have corporate enforced restrictions, largely driven by business competition. Commercial apps also often utilize data in methods that are not bound by protected health information rules. The focus on streaming data and apps also creates a loss of consumer control of their data. Once they have agreed to send streaming data to a provider, actions must be taken to stop the data sharing. This conflicts with healthcare consumers' preference to have control of what data is shared and when.

By utilizing data shared in discrete portions, consumers could remain in control of their data while gaining the insights which can be leveraged by passive monitoring, and by creating this program using open source software, it would be free of conflicting corporate interests. Such a program would need to be able to process

numerous samples automatically in order to manage the Big Data requirements of a mental health provider.

This project sought to demonstrate this model by developing a program written in the Python programming language, running on Ubuntu Linux, installed on Virtual Box Machine which would perform analyses of multiple exports of iPhone 6 Health App data. All software utilized was open source, and the data available in discrete portions. This allowed for the program to be developed without restrictions and costs inherent with commercial software and for the consumer to remain in control of their data.

When the program was run with test data, it was able to parse the iPhone xml files and generate a reference table and a time series of steps graph and a weekday mean of steps graph. The program successfully met the project objective of producing a working passive monitoring program using only open source programs and discrete data transfer, and thus demonstrated that this approach is a viable way to utilize passive monitoring while encouraging consumer control of data, avoiding interference from corporate restrictions, and being Big Data informed.

The project had various limitations related to the small scope of each particular function of the program, but further development could work on increasing the robustness of each individual module. Specific areas of development would be creating a smart parser which would use different algorithms for different iPhone versions, increasing the depth and aesthetics of the visualizations, and exploring android parsing applications. The project also sacrificed some ease of use in order to leverage the open source program, and may not be the best solution in clinicians are not familiar or willing to learn the commandline. The program also requires development in organizing the input and output files and more effectively preserving identification information, possibly using an open source database application, like MySQL.

This project was able to demonstrate how open-source programs could be paired with smartphone data exported discretely to create a program which could conduct passive monitoring while maintaining consumer control of personal data and independence from conflicting corporate interests. Continued exploration of ways to increase control of personal data and transparency is critical, especially for mental health consumers. As big data analytics and applications continue to develop, the potential for them to be misused is great. Governments, corporations, and other powerful entities have the resources to leverage control over data in ways that people may not generally be aware of or approve of.

This also can be true of mental health providers to consumers. There is a power differential, partly driven by systemic elements of the legal, medical, and mental health, but also driven by the person's need for change. People struggling with mental illness (and everyone else) often are desperate to change some circumstance in their life, and may be willing to give up something to get it. The danger for the mental health provider is to use that take away a freedom from that person for the sake of their greater good. In this context, that would be asking the person to give up their privacy in order that they can receive better services. A time comes when drastic measures may be necessary, such as when a person is a danger to themselves or others. However, if an option to preserve some of their dignity remains, it should be chosen. That is part of the motivation of this project, to provide a way to help people,

while preserving their dignity. Sometimes all it takes is to rewrite the script.

## ACKNOWLEDGMENTS

The researcher would like to thank Professor Gregor von Laszewski for his instruction and support for this project, the teaching assistants for their insight and guidance, and the anonymous participants who provided test data.

## REFERENCES

- [1] APA Practice Organization. 2017. Distinguishing Between Screening and Assessment for Mental and Behavioral Health Problems. Webpage. (2017). [www.apapracticecentral.org/reimbursement/billing/assessment-screening.aspx](http://www.apapracticecentral.org/reimbursement/billing/assessment-screening.aspx)
- [2] Apple Inc. 2017. Apple Developer Program. website. (2017). <https://developer.apple.com/programs/>
- [3] Apple Inc. 2017. Introducing ResearchKit. website. (2017). <http://researchkit.org>
- [4] Kelly Caine and Rima Hanania. 2013. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association* 20, 1 (Jan. 2013), 7–15. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsbas&AN=edsbas.fthighwire.oai.open.archive.highwire.org.amiainjl20.1.7&site=eds-live&scope=site>
- [5] Canonical Ltd. 2017. Download Ubuntu Desktop. website. (2017). <https://www.ubuntu.com/download/desktop>
- [6] C.L. Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, Supplement C (2014), 314 – 347. <https://doi.org/10.1016/j.ins.2014.01.015>
- [7] Alex E Crosby, Beth Han, LaVonne A G Ortega, Sharyn E Parks, and Joseph Gfroerer. 2011. Suicidal thoughts and behaviors among adults aged 18 years—United States, 2008–2009. *Morbidity And Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002) 60*, 13 (2011), 1 – 22. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=22012169&site=eds-live&scope=site>
- [8] Neil Elias. 2017. hid312. website. (2017). <https://github.com/bigdata-i523/hid312>
- [9] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 2 (2015), 137 – 144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- [10] Jessica D. Goodman, James R. McKay, and Dominick DePhilippis. 2013. Progress monitoring in mental health and addiction treatment: A means of improving care. *Professional Psychology, Research and Practice* 44, 4 (2013), 231. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsogo&AN=edsogl.354463723&site=eds-live&scope=site>
- [11] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43 – 49. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edselp&AN=S2352154617300384&site=eds-live&scope=site>
- [12] Diego Hidalgo-Mazzei, Andrea Murru, Mara Reinares, Eduard Vieta, and Francesc Colom. 2016. Big Data in mental health: a challenging fragmented future. *World Psychiatry: Official Journal Of The World Psychiatric Association (WPA)* 15, 2 (2016), 186 – 187. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=27265716&site=eds-live&scope=site>
- [13] Max Huang. 2017. Fix No Module Named Tkinter Issue. website. (April 2017). <http://gangmax.me/blog/2017/04/13/fix-no-module-named-tkinter-issue/>
- [14] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. 2016. Big data privacy: a technological perspective and review. *Journal of Big Data* 3, 1 (26 Nov 2016), 25. <https://doi.org/10.1186/s40537-016-0059-y>
- [15] Diego Librenza-Garcia, Bruno Jaskulski Kotzian, Jessica Yang, Benson Mwangi, Bo Cao, Luiza Nunes Pereira Lima, Mariane Bagatin Bermudez, Manuela Vianna Boeira, Flvio Kapczinski, and Ives Cavalcante Passos. 2017. The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews* 80 (2017), 538 – 554. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edselp&AN=S0149763417300337&site=eds-live&scope=site>
- [16] J.P. Medved. 2016. The Top Free and Open Source Mental Health Software. website. (April 2016). <https://blog.capterra.com/top-free-open-source-mental-health-software/>
- [17] Thomas D. Meyer, Rebecca Casarez, Satyajit S. Mohite, Nikki La Rosa, and M. Sriram Iyengar. 2018. Novel technology as platform for interventions for caregivers and individuals with severe mental health illnesses: A systematic review. *Journal of Affective Disorders* 226, Supplement C (2018), 169 – 177. <https://doi.org/10.1016/j.jad.2017.09.012>
- [18] David C. Mohr, Michelle Nicole Burns, Stephen M. Schueller, Gregory Clarke, and Michael Klinkman. 2013. Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry* 35, 4 (2013), 332 – 338. <https://doi.org/10.1016/j.genhosppsych.2013.03.008>
- [19] Scott Monteith, Tasha Glenn, John Geddes, Peter C. Whybrow, and Michael Bauer. 2016. Big data for bipolar disorder. *International Journal of Bipolar Disorders* 4, 1 (11 Apr 2016), 10. <https://doi.org/10.1186/s40345-016-0051-7>
- [20] Katherine Noyes. 2010. 10 Reasons Open Source Is Good for Business. website. (Nov. 2010). [https://www.pcworld.com/article/209891/10\\_reasons\\_open\\_source\\_is\\_good\\_for\\_business.html](https://www.pcworld.com/article/209891/10_reasons_open_source_is_good_for_business.html)
- [21] National Institute of Mental Health. 2017. (2017). <https://www.nimh.nih.gov/health/statistics/index.shtml>
- [22] Open Source Initiative. 2017. The Open Source Definition (Annotated). website. (2017). <https://opensource.org/osd-annotated>
- [23] Oracle Inc. 2017. Download VirtualBox. website. (2017). <https://www.virtualbox.org/wiki/Downloads>
- [24] Sebastien Page. 2015. How to export and import your Health data. (Jan. 2015). <http://www.idownloadblog.com/2015/06/10/how-to-export-import-health-data/>
- [25] Python Foundation. 2017. PyPI. website. (2017). <https://pypi.python.org/pypi>
- [26] Python Foundation. 2017. python. website. (2017). <https://www.python.org>
- [27] Python Foundation. 2017. Python 3.5.2. website. (2017). <https://www.python.org/download/releases/3.5.2/>
- [28] Stephen M. Schueller, Kathryn Noth Tomasoni, and David C. Mohr. 2017. Integrating Human Support Into Behavioral Intervention Technologies: The Efficiency Model of Support. *Clinical Psychology: Science and Practice* 24, 1 (2017), 27. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsqao&AN=edsqcl.487768361&site=eds-live&scope=site>
- [29] Social Security Administration. 2017. (2017). [https://www.ssa.gov/policy/docs/statcomps/di\\_asr/2013/di\\_asr13.pdf](https://www.ssa.gov/policy/docs/statcomps/di_asr/2013/di_asr13.pdf)
- [30] Z. Steel, C. Marnane, C. Iranpour, Tien Chey, J. W. Jackson, Patel Vikram, and D. Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology* 43, 2 (2014), 476 – 493. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=llhh&AN=20143278163&site=eds-live&scope=site>
- [31] Substance Abuse and Mental Health Services Administration. 2015. (2015). [https://www.samhsa.gov/data/sites/default/files/2015\\_National\\_Mental\\_Health\\_Services\\_Survey.pdf](https://www.samhsa.gov/data/sites/default/files/2015_National_Mental_Health_Services_Survey.pdf)
- [32] Substance Abuse and Mental Health Services Administration. 2017. Behavioral Health Treatments and Services. (2017). <https://www.samhsa.gov/treatment>
- [33] Vincent Tournaire. 2016. How to setup a ResearchKit project. website. (Feb. 2016). <http://blog.shazino.com/articles/dev/researchkit-setup-project/>
- [34] Tori Utley. 2017. Could This New ResearchKit App Help Develop The Fitness Tracker For The Brain? website. (May 2017). <https://www.forbes.com/sites/toriutley/2017/05/31/could-this-new-researchkit-app-help-develop-the-fitness-tracker-of-the-brain/#4a8848677c9e>
- [35] Gregor von Laszewski. 2016. 4.3. Python. website. (2016). <https://cloudmesh.github.io/classes/1523/2017/python.html>
- [36] Sam Stephenson Yuu Yamashita. 2017. Simple Python Version Management: pyenv. website. (2017). <https://github.com/pyenv/pyenv>

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
```

```
=====
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-16 09.36.28] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Typesetting of "report.tex" completed in 1.1s.
```

```
=====
Compliance Report
```

```
=====
name: Neil Eliason
hid: 312
paper1: Review on 3 Nov 2017
paper2: 100%
project: 99%
```

```
yamlcheck
```

wordcount

---

8

```
wc 312 project 8 6221 report.tex
wc 312 project 8 7072 report.pdf
wc 312 project 8 2582 report.bib
```

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

passed: False

find input{format/final}

---

passed: False

floats

---

```
147: \begin{table}[htb]
149: \includegraphics[width=\columnwidth]{images/Client1.pdf}
154: \begin{figure}[htb]
155: \includegraphics[width=1.0\columnwidth]{images/Client1StepsByDate
    .pdf}
159: \begin{figure}[htb]
160: \includegraphics[width=1.0\columnwidth]{images/Client1MeanStepsBy
    DayOfWeek.pdf}
164: \begin{figure}[htb]
165: \includegraphics[width=1.0\columnwidth]{images/Client1StDevStepsB
    yDayOfWeek.pdf}
169: \begin{figure}[htb]
```

```
170: \includegraphics[width=1.0\columnwidth]{images/Client1Scatterplot  
OfStepsByDayOfWeek.pdf}
```

```
figures 4  
tables 1  
includegraphics 5  
labels 0  
refs 0  
floats 5
```

```
True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
False : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check  
passed: True
```

```
When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction
```

---

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# The Impact of Clinical Trial Results on Pharmaceutical Stock Performance

Tiffany Fabianac

Indiana University

Bloomington, Indiana 47408, USA

tifabi@iu.edu

## ABSTRACT

While many relate stock market trading to gambling, successful traders have turned stock picking into a science. The likes of Warren Buffet tell us that successful stock buying is all in the research. So what kind of research aids in the prediction of companies within the highly volatile pharmaceutical market? The use of available, open-source APIs and Google Alerts are used to explore if clinical trial results can directly impact stock performance in small, mid, and large cap pharmaceutical companies. Key words and/or phrases in results and related news articles are identified as possible predictors of market effect. As well as a comparison to already established analyst ratings from Barclays, Goldman, and J P Morgan Chase which have already been shown to impact stock performance.

## KEYWORDS

Big Data, HID313, i523, Stock Market, Pharmaceutical

## 1 INTRODUCTION

A “stock” is a piece of ownership in a company. Offering stocks for sale provides capital to the selling company in exchange for a stake in the company. A stock market is a collection of exchanges where trading of stocks takes place [13]. Evidence of early stock markets date back to the fourteenth century with the offering of state loan stocks throughout Italy. Even prior to the organization of stock markets, price fluctuations for goods such as wheat and barley were tracked by early economists. The first “modern” stock market appeared in Amsterdam in the seventeenth century where the volume of stocks traded and the fluidity in which they were traded reached a new high [4].

The biggest stock markets in the world are currently the New York Stock Exchange (NYSE), the National Association of Securities Dealers Automated Quotations (NASDAQ), and the London Stock Exchange. NYSE started in 1792 with twenty four stock brokers. The initial focus was government bonds which provided secure, long term income. The early days of the 1800 saw stocks traded through telegraph. Telephones replaced the telegraphs in 1878. Current trading on the NYSE can surpass 1.4 billion shares each day across almost 4,000 companies [8]. NASDAQ began as an all-electric equities exchange in 1971 and today provides trading, technology, and information services for financial markets. Today over 4,000 companies are traded on the NASDAQ with over 1.8 billion trades per day [20]. The London stock exchange was founded in 1801. Currently over 2,600 companies across 60 countries are traded on the London Stock Exchange each day [6].

Throughout the history of markets, prices have been tracked and insightful traders have attempted to predict and capitalize on price fluctuation. The age of computers opened new doors for stock

analysis and trend prediction to facilitate capital gains for traders. Financial companies like Goldman Sachs and JPMorgan Chase & Co. have hired mathematicians, statisticians, and trade analysts since the early days of trading in an effort to predict the market in a consistent manner. Once an algorithm is established and used consistently the algorithm itself but be considered as a variable that could effect the prediction outcome [? ].

A major complexity in creating algorithms for the stock market is that the market tends to follow the erratic emotions and feelings of humans. If computers were running the market, making trade decisions based on logic and reason, then the market would be much more stable. The volatility of human emotions about money and stocks creates tremendous volatility in the market. The revolution of social media has provided a means of measuring the mood of possible traders. For this reason, the ability to predict society’s reaction to news has developed into a field of study within the data science world [3].

How big of an impact can news articles have on the stock market? In September 2008, an article published on a South Florida News website reported that United airlines had filed Chapter 11 bankruptcy. The news struck so hard that United’s stock plummeted 75% from \$12 to \$3. Interestingly enough, the article was just about six years old and had originally been published by the Chicago Tribune in December 2002. Even though the report was literally “old news” it did not prevent massive panic from investors [28].

### 1.1 Pharmaceutical Sector

The pharmaceutical industry has evolved around the need to establish drugs and treatment options for diseases. Research and development within pharmaceutical companies range from compound identification to disease characterization. This market is directly affected by the results of drug tests such as clinical trials and the establishment of new treatment options. Market growth also comes from manufacturing and licensing of drugs and treatment methods. Innovation is the key driver of this industry [9].

Like the financial sector trying to predict the stock market, the pharmaceutical industry has devoted resources to developing prediction algorithms and machine learning systems. The efforts of drug manufacturers are aimed to create a system that consistently predicts or aids in identifying drug targets. One such approach is the development of virtual screening for drug discovery meant to reduce the experimental failures associated with high throughput screening. High throughput screening is carried out to test many chemicals, molecules, compounds, proteins, hormones, viral vectors, etc all at once on large grids or plates which can test many different treatment combinations all together. Large costs and big

data sets are associated with high throughput screening which is now becoming virtual with the help of advanced molecular profiling [14].

## 1.2 Clinical Trials

A clinical trial is a planned experiment involving patients with the intent to elucidate an appropriate or effective treatment option(s) for the population of patients afflicted with the same medical condition. A big concern with clinical trials is that inferences are made for the entire population of patients from a relatively small sample size [22]. One of the first clinical trials recorded was carried out in the eighteenth century to evaluate six treatments on twelve patients with scurvy. Two patients that were given oranges and lemons recovered very quickly. Fisher introduced the concept of randomization in the nineteenth century [7].

Clinical trials have four defined phases. Phase I trials identify how well a drug is tolerated by determining the maximally tolerated dose (MTD) on a very small sample size. Phase I trials have very simple experimental designs as the only intent is to examine toxicity. Phase II explores biological activity or effect on a small patient sample size. The design of a Phase II trial is dependent on the design on the Phase I trial as both share the intent to evaluate adverse events. Phase III trials follow the design of Phase II trials but on a bigger sample size with the intent to solidify a treatment's effectiveness in clinical practice. Phase IV trials are prolonged Phase III trials that can track a drug, procedure, or instrument for decades with continuous efficiency reflection [7].

Clinical trial designs have been very slow to evolve due to restrictions enforced by governing agencies such as the US Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC). While these restrictions are intended to minimize patient risk, they also greatly restrict the potential of clinical trial data collection. Other limiting factors include difficulty enrolling high quality participants for each trial phase, problems monitoring how well patients are following protocol, difficulty sorting out "the placebo effect" or the ability for patients to feel as if they are recovering without actually receiving treatment, and overall minimizing poor quality of data [7].

## 1.3 Established Analyst Ratings

Companies within the financial sector often publish rankings of the top stocks that the company invests in. The ratings are a way to attract investors with proof that the company is diligently analyzing the market and "picking winners". These published rankings have been shown to boost or deflate rallies behind particular stocks that are added or removed for these prestigious lists [1].

The Goldman Sachs Group, Inc. was founded in 1869. The company provides a full stack portfolio of banking and investment services. Goldman Sachs career website states that the company is driven to achieve superior returns for their clients which include pension funds, hedge funds, and mutual funds. The company boasts that their research analysts are curious and creative [26]. Goldman Sachs Global Investor Research group provides stock ratings on a scale of Buy, Neutral, and Sell [25].

J P Morgan Chase (JPM) is one of the largest investment banks in the world [27]. The company's investment mechanisms include currency, emerging markets, equities, and fixed income. JPM publishes quarterly market insight reports with "buy" and "sell" ratings for the companies of interest to the firm. Subscribers to JPM's services can even get an audio version of the report which details market trends [18].

Barclays was founded in London in 1896. The bank currently serves over forty-eight million customers and releases stock picks every quarter but for a limited number of stocks [27]. Because Barclays is so selective with their stock promotions, only selecting some 50 stocks to support, it is possible that they have a greater impact on the market than other companies in the stock prediction game.

## 1.4 Data Resources

An Application Programming Interface (API) acts as the middleman between the requesting service and the performing service. When a user or system submits a request the request is passed to the API which translates it for the processing system then returns the results in a receivable format. This project uses the free Gmail API provided by Google to read and extract data from specific email messages.

Machine learning is the study using computer language to recognize patterns and make data-driven decisions based off of them. It is based on the theories of statistics. Bayes' Theorem gives the probability of an event occurring given some evidence. Bayes' Theorem is vital in Machine Learning because it provides evidence to how probabilities should be updated given new evidence. Markov's theory describes properties that can be predicted based only on past events. Some of the first learning programs were designed to play boardgames such as checkers and chess [30].

NASDAQ's website provides historical stock performance data that can be exported as a Comma-Separated Values (CSV) file. The disadvantage of NASDAQ's free export service is that each stock must be exported separately. The free quote service can be accessed at [21]. NASDAQ provides API services for subscribers starting at \$5,000 per year [19]. Access to NASDAQ's API services can also be granted through corporate sponsorship. NASDAQ's free CSV export services were used to collect initial project data. In this example, the stock history for Celsion Corporation during the week of August 21, 2017 is shown in 1.

**Table 1: NASDAQ CSV file format example**

date	close	volume	open	high	low
2017/08/25	1.3700	179097.0000	1.3600	1.4100	1.3000
2017/08/24	1.3600	149832.0000	1.3100	1.3600	1.2810
2017/08/23	1.3100	223451.0000	1.2500	1.3300	1.2430
2017/08/22	1.2800	164594.0000	1.3200	1.3200	1.2400
2017/08/21	1.3300	169037.0000	1.3300	1.3700	1.2800

Exports such as this one offered by NASDAQ and API interfaces for stock data are provided by numerous companies. The Yahoo! Finance API is explored below and the Google Finance API was used to perform the stock data extraction for the analysis presented.

Additional resources such as stock tracking apps and free exports are available. CSV exports such as the one listed above can be downloaded from Google Finance, Yahoo! Finance, and many others. This publication does not provide a complete list of available resources, but attempts to present a few for comparison.

Python.org provides a python module to pull stock data from Yahoo! Finance [23]. The package can be installed through Git by cloning the Git directory where the package is available: [17]. To install the python package without Git the tape archive can be downloaded from [24]. Tape archives allow for compression of multiple files which can be restored to their original format using the tar command in the command line [15]. Apply the tar options: z - filter archive through gzip, x - extract an archive file, and f - filename of archive, use “cd” to change the current working directory, and then install the python module using the package management command “pip”:

```
tar -zxf yahoo-finance-1.4.0.tar.gz
cd yahoo-finance
pip install yahoo-finance
```

While Yahoo! Finance is a great resource, the API does not function consistently, and as of this writing the API has been turned off by Yahoo!.

## 2 METHODS

### 2.1 Data Collection

Data collection was initiated with the use of Google Alerts. Google allows for alerts to be configured from Google [11]. Gmail users can configure these alerts to be sent through email when news or other types of articles pertaining to a defined subject are released to the web. The Google Alerts for this project were: “Phase III Trial”, “Phase 3 Trial”, and “Meets Primary End Point”. When these phrases are detected by Google, the link to the webpage and a short description are sent via email to the configured email address. On busy days, an excess of 100 alerts were received for these alert phrases. On slow days, only a couple alerts were received. Only very infrequently were no messages received.

To collect data from the received Google Alerts without too much manual clicking, Gmail has an available API which allows users to pull data from a Gmail account. To start using the Gmail API, a user must first configure their Authentication credentials through Google’s developer console. The JSON format is shown in Table 2. Once credentials are received in the form of a JSON file, the

**Table 2: Google Gmail API JSON format**

```
{"installed":{"client_id": "###.apps.googleusercontent.com",
"project_id": "###",
"auth_uri": "[URL]",
"token_uri": "[URL]",
"auth_provider_x509_cert_url": "[URL",
"client_secret": "###",
"redirect_uris": ["urn:ietf:wg:oauth:2.0:oob",
"http://localhost"]}}
```

Google Client Library can be installed using pip to install google-api-python-client. The Google Development team has provided a quickstart file which facilitates the first authentication run. Running this quick start guide will open a browser window and prompt the user to log into a Gmail account. The user then accepts the authorization and can run the Gmail API from command line or other compilers.

Headlines of the received alerts, usually the title of the article and the first couple of lines, are referred to as “Snippets” by Google’s Gmail API. This project pulled only the Snippets and the date from the Google Alerts. The Snippets do not contain the whole article but may still provide enough evidence of sentiment for further analysis and prediction of the associated stock. Unfortunately, no solution was identified for extracting the appropriate stock symbols from the Snippets so this task had to be performed manually.

The google-api-python-client provides a number of helpful modules that are designed to provide simple access to Google APIs. The main components of authenticating the API are apiclient which build the credential string which will be added to each execution string for the API. Auth2client provides the authentication library [10]. Access to HTTP connections are provided by httplib2 [12]. Dates are managed and manipulated with time, dateutil, and datetime. Csv, io, and json provide text and file parses and manipulators.

The Python code calls the Gmail API and writes a .csv from the data. After calling all needed libraries, the scope of the authorization is defined. Google mail can be opened with a Readonly or Modify authentication. Next, the credentials are established by the JSON file received during the API authentication setup. This JSON must be saved in the same directory as the code being run. The code sets the variables for User ID and Label then runs an execution command calling the Messages.List API, which looks like this:

```
GMAIL.users().messages().list(userId='me', labelIds=
[INBOX], q='from:[ALERTS] before:[DATE]').execute()
```

Google has defined the user ID “me” as the global for the authenticated account in use. The label ID “INBOX” designates that the messages will be pulled from the inbox folder, but any other folder could be called here as well as a collection of labels that Google has defined such as “UNREAD”. The “q” designates a query. The query will return only messages from the Google Alerts email address which have been received by the twenty-fourth of November 2017. This data was selected so that all returned records would have five market days of stock prices to compare. This execution returns a dictionary which contains message IDs for all the messages that matched the query.

The next step is to “get” the messages with the use of the Messages.Get API. While looping through the dictionary of message ID from the defined query, the script retrieves the Date and Snippet for each. Additional options could return the Sender, Receiving Email, Email body, among others. The syntax is shown here:

```
GMAIL.users().messages().get(userId='me',
id=m_id).execute()
```

The user ID is the same as described previously with the ID being the current message ID within the loop. This execute command returns a dictionary which is parsed from “payload” to “headers” to extract the Date. The Snippet is also grabbed from the message

dictionary and along with the Date, passed to a final list to be written to a .csv file.

Figure 1 shows the entire code to extract Google Alerts data using the Google provided Gmail API.

The Python package pandas is an incredible resource that provides a number of tools to read, parse, extract, and manipulate delimited file or data types. The Pandas package has a resource for getting stock market data from free online sources such as Yahoo! mentioned above and Google. To install this package through Git, simply clone the directory, use the “Change Directory” command “cd” to change the current working directory, and installing the python module as follows: “ git clone git://github.com/pydata/pandas-datareader.git cd pandas-datareader python setup.py install ”

If the Python setup returns the error: “python: command not found” run the following with the path to the python installation:

```
“PATH=$PATH:/c/Python27”
```

Pandas-datareader and many other packages can also be installed via pip. In example, many additional packages are needed to run a python script using pandas-datareader. These packages can be configured all at once or one at a time as follows: “ pip -m install -user numpy scipy matplotlib ipython jupyter pandas sympy nose urllib3 chardet idna ”

Unlike the NASDAQ export, using Google as a data source for pandas-datareader requires each attribute to be called separately. This means calling the Close Price, Open Price, High Price, etc individually and joining them through code. Also, unlike NASDAQ’s export but this time in a positive light, multiple tickers can be passed together. This allows for all historical data to be pulled for many stocks with a single code.

The Python code for collecting historical stock data is propelled by pandas\_datareader. The script starts by reading in the .csv created using the Google API script described previously. The data is read in as a dictionary using DictReader and the output file is opened/created right afterwards to allow for writing out with each loop through the starting file’s dictionary. For each line the stock ticker and date of the Google Alert are passed to a function that returns the highest price of the stock 5 days after the Google Alert, the stock and ticker are then passed to a function that pulls the opening price on the day that the Google Alert was received. The highest price and starting price are the used to calculate the percent change using the formula: “ round(((high-startPrice)/startPrice) × 100,2) ” If the high price is 10% higher than the starting price the line is given a “W” for “Winner”. If the high price is less than 10% of the starting price then the line is marked with a “L” for loser. The whole line with the addition of the Win or Lose designation and the percent change is written to a new .csv file with the intention of attempting sentiment analysis with the Win or Lose designations as the outcome and the Snippets as the sentiment.

Figure 2 shows a portion of the code to combine the data produced by the Google Alert mining and available historic stock price data.

Twelve out of over one hundred stock tickers returned by Google Alerts were flagged at “Winners” for increasing in price by 10% within five days after the Google Alert was received and are shown in Table 3.

ABEO Snippet appears to reflect a number of disappointments followed by something positive: “ Abeona Therapeutics - String Of

**Table 3: Winning Stock Tickers**

Ticker	prctChange	High	Open	Date
['ABEO']	27.39	10.0	7.85	2017-08-22
['ARRY']	15.41	10.11	8.76	2017-08-22
['CLSN']	160.9	3.47	1.33	2017-11-23
['EARS']	20.83	0.87	0.72	2017-11-18
['EGLT']	15.04	1.3	1.13	2017-11-17
['HCM']	39.87	35.01	25.03	2017-11-19
['NLNK']	57.8	10.02	6.35	2017-11-18
['NWBO']	45.0	0.29	0.2	2017-11-19
['NWBO']	45.0	0.29	0.2	2017-11-17
['ONCE']	11.53	83.19	74.59	2017-08-21
['OTIC']	11.47	20.9	18.75	2017-08-23
['PSTI']	32.23	1.6	1.21	2017-11-22
['VTVT']	10.92	5.08	4.58	2017-11-23
['VTVT']	24.24	5.69	4.58	2017-11-19
['VTVT']	24.24	5.69	4.58	2017-11-18

Pearls Strategy With Numerous Catalysts And A Lot Of Upside ” This Snippet was received August 22, when ABEO’s stock opened at \$7.85. The stock hit its five year high of \$19.95 on October 10.

ARRY is a bio-pharmaceutical company that was call out in the training set as a “winner” for August 22. J P Morgan Chase & Co confirmed a “buy” rating for ARRY on September 11, three weeks after it was identified by this model as a “winner”. Goldman Sachs increase their buy in to ARRY on October 22 by 33%. The Snippet for ARRY does not appear to reflect a positive sentiment about the company: “ Array Biopharma (ARRY) Reaches \$8.58 After 7.00% Down Move; Per Se Technologies ”

CLSN started the year just under \$10 a share and slowly declined to its current \$2.40. The Snippet for CLSN was received on November 23 when the stock briefly rose 160% before falling again: “ After Reaching Milestone, Is Celso Corporation (NASDAQ:CLSN)’s Short Interest Revealing ”

EARS is a small tier stock with a market cap of \$19 million. The stock rose to \$0.93 per share on November 24 before falling to \$0.42 on November 28. The Snippet depicts analysts predictions of negative earnings: “ Analysts See \$-0.20 EPS for Auris Medical Holding AG (EARS) BZ Weekly The Company’s advanced product candidate, AM-101, is in ”

Egalet Corporation (EGLT) develops abuse resistant formulations of opioids. The Snippet is overwhelming positive and describing stock increases: “ Egalet progressing second abuse-deterring opioid med The Pharma Letter Egalet (Nasdaq: EGLT) says its share move up a hefty 38.55% ” This Google Alert was receive on November 17, just prior to another 30% stock increase.

HCM is a pharmaceutical company headquartered in China. The Snippet reflects the companies one year growth of over 160%: “ Will Hutchison China MediTech Limited (HCM) Run Out of Steam Soon? BZ Weekly ... Hutchison China MediTech Limited (LON:HCM) were ”

NLNK received positive feedback from established analysts on November 18. Causing the stock to briefly rise and then return. This Snippet and change may reflect the power on analyst ratings: “

```

...
Portion of Reading GMAIL Alerts using Python
Tiffany Fabianac Modified code from:
- https://github.com/abhishekchhibber/Gmail-Api-through-Python
- Abhishek Chhibber
...
# Creating a storage.JSON file with authentication details
SCOPES = 'https://www.googleapis.com/auth/gmail.modify'
store = file.Storage('storage.json')
creds = store.get()
if not creds or creds.invalid:
    flow = client.flow_from_clientsecrets('client_secret.json', SCOPES)
    creds = tools.run_flow(flow, store)
GMAIL = discovery.build('gmail', 'v1', http=creds.authorize(Http()))

user_id = 'me'
label_id_one = 'INBOX'

alert_msgs = GMAIL.users().messages().list(userId='me', labelIds=[label_id_one],
    q='from:googlealerts-noreply@google.com').execute()

# We get a dictionary. Now reading values for the key 'messages'
mssg_list = alert_msgs['messages']

final_list = []

for mssg in mssg_list:
    temp_dict = {}
    m_id = mssg['id'] # get id of individual message
    message = GMAIL.users().messages().get(userId=user_id, id=m_id).execute() # fetch the message using API
    payld = message['payload'] # get payload of the message
    headr = payld['headers'] # get header of the payload

    for two in headr: # getting the date
        if two['name'] == 'Date':
            msg_date = two['value']
            date_parse = (parser.parse(msg_date))
            m_date = (date_parse.date())
            temp_dict['Date'] = str(m_date)
        else:
            pass

    temp_dict['Snippet'] = message['snippet']

    final_list=json.dumps(temp_dict) # This will create a dictionary item in the final list
    re.sub(r'\u22c5', '', final_list)

```

**Figure 1: The Google API Python code calls the Gmail APIs Messages.list which lists reduced properties of Gmail messages and Messages. Get which returns the messages themselves. Lists is used to query the messages that are wanted based on the defined criteria: userId=me, labelIds=INBOX], q=from:googlealerts-noreply@google.com. Get then retrieves the messages identified in using List and returns the messages content for Date and Snippet.**

NewLink Genetics Corporation (NASDAQ:NLNK) Given Buy Rating at Cantor Fitzgerald StockNewsTimes Indoximod is expected to enter a ”

NWBO held steady through October at \$0.16 and between until November 15 and November 28 rose 87%. The Snippet was received on November 18 in the prime of the increase. “ Here’s Why Northwest Biotherapeutics, Inc (OTCMKTS:NWBO) Just Ripped Higher

```

...
Collect Historical Stock Data
Tiffany Fabianac Modified code from:
- http://pandas-datareader.readthedocs.io/en/latest/remote_data.html
...
def stockData (startDate, endDate, ticker):
    # Define which online source one should use
    data_source = 'google'

    # Use pandas_reader.data.DataReader to load the desired data.
    panel_data = data.DataReader(ticker, data_source, startDate, endDate)

    close = panel_data.ix['Close']
    volume = panel_data.ix['Volume']
    op = panel_data.ix['Open']
    high = panel_data.ix['High']
    low = panel_data.ix['Low']

    # Getting all weekdays between 01/01/2017 and 12/31/2017
    all_weekdays = pd.date_range(start=startDate, end=endDate, freq='B')

    # Align new set of dates
    close = close.reindex(all_weekdays)
    volume = volume.reindex(all_weekdays)
    op = op.reindex(all_weekdays)
    high = high.reindex(all_weekdays)
    low = low.reindex(all_weekdays)

    result = pd.concat([close, volume, op, high, low], axis=1, join='inner')
    result.columns=['close','volume','open','high','low']
    return result

def findHigh (startDate, ticker):
    # Get date and five days after
    temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")
    endDate = temp_date + BDay(5)

    result = stockData(startDate, endDate, ticker)
    tempHigh = result.nlargest(1,'high')
    high = tempHigh.iloc[0]['high']
    return high

def openPrice (startDate, ticker):
    temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")
    endDate = temp_date + BDay(1)

    result = stockData(startDate, endDate, ticker)
    open = result.iloc[0]['open']
    return open

```

Figure 2: This Python script takes in the Date, Stock Ticker Symbol, and Snippet from the Google API.csv that was produced using both manual mining of the stock symbols and the python script provided for getting the Date and Snippet from Gmail. This code returns a modified .csv which lists an "L" for stocks that did not increase by 10% in five days and a "W" for stocks that increased by at least 10%. It also prints the stocks that increased by at least 10% along with the highest price over 5 days, the starting price on the day that the Google Alert was received, and the percent change.

The Finance Registrar The Company's lead program orthwest Bioth Cmn (NASDAQ:NWBO) Stock fi? Is it Overbought? First News 24 The Business's lead product, DCVax-L, is in an ongoing "

ONCE is a large cap therapeutics company which showed growth through September. The Snippet reflects news of a changed analyst rating: " Spark Therapeutics Inc (ONCE) is Initiated by Evercore ISI to "In-line" "

OTIC rose in August just before crashing from \$20.18 to \$3.20 after a failed Phase III clinical trial in September. The Snippet captured analyst confidence in the company: " Otonomy (OTIC): Reiterating Outperform Ahead Of Catalysts - Cowen "

PSTI is a leading developer of cell therapy products derived from placenta. The Snippet received on November 22 reflects news of a granted patent application: " Pluristem Therapeutics (PSTI) Granted US Patent for Skeletal Muscle Regeneration StreetInsider.com This very important patent comes"

VTVT's Snippets reflect stock decreases, low sentiment scores, and drug treatment competition: " vTv Therapeutics (VTVT) Reaches \$5.01 After 5.00% Down Move; FMC (FMC) Shorts Down By vTv Therapeutics (VTVT) Receives Media Sentiment Rating of 0.25 The Lincolnian Online vTv Therapeutics Inc is a clinical-stage Head-To-Head Comparison: vTv Therapeutics (VTVT) versus Its Competitors The Ledger Gazette Its drug candidate for the treatment of " These sentiments do not reflect positive news and should be cause to look more deeply at the stock comparison being performed.

## 2.2 Data Analysis

There are many methods for analysis that could be implemented for this dataset. Time series prediction could be used to identify trends in the stocks of interest [2]. Regression analysis is very common to identify key factors that contribute to the accuracy of a prediction. TextBlob sentiment analysis allows for sentiment analysis to be performed in as little as four lines of code. TextBlob returns a number between -1 and 1 for how negative (-1) or positive (1) a defined sentiment or group of text is [16]. Tensorflow is another popular way of creating sentiment analysis which takes an input of words with the intent of returning a sentiment of positive, negative, or neutral. In order to do this Tensorflow uses a build in learning and training set called tflearn to compare previously established sentiments. For example, words like "love" and "happy" return a positive sentiment while words like "hate" and "sad" return a negative sentiment [5].

Random Forest algorithms create decision trees for each variable. Each tree represents the sequence of events or decisions that led to the outcome or result. With each branch or step through the decision tree a probability is calculated for the outcome and the collection of trees work are combined to create multiple "regression lines" that are used to predict an outcome when presented with new data that does not have an outcome. The model or collection of trees form what is called a random forest can then be used to predict sentiment or outcomes. For stock data or other time series datasets, it is essential to continuously re-train the model to perform at its best. As mentioned above it is possible for additional models and even the model itself to begin to influence the prediction model.

The code that performs random tree analysis starts with some dependencies. Os is imported to allow for command line functionality,

the machine learning library sklearn is used because it has a very fast learning rate, KaggleWord2VecUtility is a utility that processes raw text into segments for learning, pandas as mentioned before helps with delimited file manipulation, nltk that already contains a number of words and phrases that are not useful for sentiment analysis importing this library helps to eliminate those elements from the dataset we are training on. To install KaggleWord2VecUtility visit the DeepLearningMovies github directory [29].

In this code the Kaggle module removes special characters associated with HTML. It was intended to return a URL from the Google Alerts and run the website associates with each alert through beautiful-soup to use the entire article as training data, but the Gmail messages were encoded in such a way that it was not possible to extract the URL from the Google Alert. Nltk removes words such as "to" or "the" which do not hold any inherent meaning that could be applied to the sentiment analysis. The cleaning process converts the first Snippet as follows: " Abeona Therapeutics - String Of Pearls Strategy With Numerous Catalysts And A Lot Of Upside abeona therapeutics string pearls strategy numerous catalysts lot upside "

Once the Snippets are free of special characters and non-sentiment words, they are parsed into a vector. This process creates what is called a "Bag of Words" by creating a dictionary with the count of each word in the text. This is also called tokenization or vectorizing and is performed easily with the sklearn package's countVectorizer process. Here the analyzer is set to word, there is no defined tokenizer, pre-processor, or stop words needed so these are set to "None". The maximum number of features controls the limit on the maximum number of words and frequencies contained in the bag of words.

A model is easily created from the defined bag of words using sklearn's fit\_transform which is converted to an array. The method for classification is a random forest which builds decision trees for each variable in the dataset. In example, the first Snippet describes a "winning" variable and contains the word "Upside" if other Snippets contain the word "Upside" it might be indicative of a "winning" classifier. The last step calculates predictions for the new dataset based on the established classifiers. This is simple done with the RandomForestClassifier's predict function.

Figure 3 shows the entire code to train on the dataset provided by the historical stock data and Google Alert sentiments.

The Python code for verifying the random tree analysis by pulling historical stock data for each ticker analyzed is propelled by pandas\_datareader. The script starts by reading in the .csv created using the random tree analysis script described previously. The data is read in as a dictionary using DictReader and the output file is opened/created right afterwards to allow for writing out with each loop through the starting file's dictionary. For each line the stock ticker and date are passed to a function that returns the highest price of the stock from the date of the received alert to the current date, the stock and ticker are then passed to a function that pulls the opening price on the day that the Google Alert was received. These two prices are compared to verify if the stock increased by 10% from the time of the alert.

Figure 4 shows a portion of the code to combine the data produced by the random forest analysis and combine it with available historic stock price data.

```

...
PORTION OF: Use KaggleWord2VecUtility to produce random forest analysis
Tiffany Fabianac Modified code from:
- https://youtu.be/AJVP96tAWxw
- Siraj Raval
...

# Cleaning and parsing the training set
for i in xrange(0, len(train['Snippit'])):
    clean_train_reviews.append(" ".join(KaggleWord2VecUtility.
review_to_wordlist(train['Snippit'][i], True)))
#Creating the bag of words
vectorizer = CountVectorizer(analyzer="word", tokenizer=None, preprocessor=None,
stop_words=None, max_features=5000)
train_data_features = vectorizer.fit_transform(clean_train_reviews)
train_data_features = train_data_features.toarray()

#Training Random forest
forest = RandomForestClassifier(n_estimators=100)
forest = forest.fit(train_data_features, train['W/L?'])
clean_test_reviews=[]

#"Cleaning and parsing
for i in xrange(0,len(test['Snippit'])):
    clean_test_reviews.append(" ".join(KaggleWord2VecUtility.
review_to_wordlist(test['Snippit'][i], True)))
test_data_features = vectorizer.transform(clean_test_reviews)
test_data_features = test_data_features.toarray()

print "Predicting test labels...\n"
result = forest.predict(test_data_features)

```

**Figure 3: The Sentiment Python code takes the .csv exported by the historical stock script and parses the Snippets to train on the stock script and apply it to more recent stock quotes and Google Alerts**

### 3 RESULTS

The resulting CSV file contains the accuracy of the prediction, if the stock did not increase by atleast 10%, the date that the Google Alert was received, the Sentiment that was calculated by the random forest algorithm, and the stock ticker. The results export to a .csv as shown in Tables 4, 5, and 6.

This analysis shows the stock ticker ABBV for the pharmaceutical company AbbVie as a “loser” twice as two alerts were received about the company on December 3 and 4. As of December 4 ABBV is down 1.08% post Google Alert receipt. ACAD is the ticker for ACADIA Pharmaceuticals Inc. which is down 1.09% since receipt of the Google Alert on November 30. Alnylam Pharmaceuticals, Inc (ALNY) is down 1.06% since December 2. ARGX is down 0.97% since receipt of the Google Alert but up over 18% for the prior five days. ARGX did not appear in the training data set so it might be worth while to explore factors that contributed to it’s recent increase, if not clinical trials. Interestingly, BABA is a Chinese e-commerce site which is down 2.88%. This ticker appearing is cause to look closer at the article that was link to the Clinical Trial Alert but returned a retail chain.

#### 3.1 Comparison to Established Analyst Ratings

One of the important aspects of professional analyst ratings is that the intent is to identify the best long term investments. This project only looked at short term success over a period of five days. Further research should refine additional models to compare success in shorter term, one day, and longer term, six months to a year or more.

ABBV, a predicted “losers”, is marked “Neutral” by JPM. The two Snippets stored for ABBV are: “ Cornercap Investment Counsel Has Raised Abbvie Com (ABBV) Stake; Profile of 7 Analysts ... NormanObserver.com The firm also develops AbbVie Inc. (NYSE:ABBV) Updates On Phase III Murano Trial MMJ Reporter AbbVie Inc. (NYSE:ABBV) reported that the American Society of ” The ABBV Snippets do not appear to be negative, and may even swing more in the positive light. AbbVie being a large cap pharmaceutical company may create lower volatility for the stock. Reanalyzing the data and splitting companies into small, mid, and high tier categories may give very different results over long term and short term growth. Larger companies, with many more investors, tend to be more stable.

ACAD, a predicted “losers”, is marked as “Neutral by Goldman Sachs. Interestingly enough, the Snippet about ACAD mentions a

```

...
PORTION OF: Validate random forest analysis
Tiffany Fabianac Modified code from:
- http://pandas-datareader.readthedocs.io/en/latest/remote_data.html
...
with open('randomForestResults.csv', 'rb') as csvfile:
    with open('resultsData.csv','wb') as f:
        datareader = csv.DictReader(csvfile)
        writer = csv.DictWriter(f, fieldnames=datareader.fieldnames, extrasaction='ignore',
                               delimiter=',', skipinitialspace=True)
        writer.writeheader()
        for line in datareader:
            if (line['Ticker'] == '' or line['Sentiment'] == ''):
                pass
            else:
                ticker = [line['Ticker']]
                date = line['Date']
                high = findHigh(date, ticker)
                startPrice = openPrice(date, ticker)
                prctIncrease = round(((high-startPrice)/startPrice)*100,2)
                if (high > startPrice*1.1 and line['Sentiment']=='W' ):
                    line['Accuracy']='W'
                    print ticker, prctIncrease, high, startPrice, date
                else:
                    line['Accuracy']='L'
                writer.writerow(line)

```

**Figure 4:** This Python script takes in the Date and Stock Ticker Symbol from the sentiment .csv that was produced using the sentiment python script provided for performing a random forest analysis on the Google Alert results. This code returns a modified .csv which lists an “L” for stocks that did not increase by 10% from the time the Alert was received to the current date and a “W” for stocks that increased by at least 10%. It also prints the stocks that increased by at least 10% and were marked as “winners” by the sentiment script.

sentiment ranking which is actually what would be considered a positive rating: “ EPS for The Kroger Co. (KR) Expected At \$0.41; Acadia Pharmaceuticals (ACAD)s Sentiment Is 1.05 San Times The Company ” Increasing the Google Alert scope to include data related to sentiment for pharmaceutical companies may be beneficial to the model.

ALNY, a predicted “losers”, is marked as “Buy” by JPM, Goldman Sachs, and Barclays. These analyst ratings may indicate that the model is not a good indicator of long term success as the analyst ratings suggest. This requires greater research which should include increasing the historic interval from five days to six months or more. The Snippet does not seem to reflect anything positive or negative about the company: “ How Analysts Rated Alnylam Pharmaceuticals Inc. (NASDAQ:ALNY) Last Week BZ Weekly The company’s clinical development programs ” ARGX, a predicted “losers”, is ranked as “Underweight” by Barclay, as recently upgraded to “Buy” by Goldman Sachs, and has been downgraded to “Neutral” by JPM. The Snippet used to rate this company mentions a number of other stock tickers but gives the impression that ARGX should be a stock of interest for would be investors: “ Here’s Why You Need To Keep An Eye On ARGX MGNX KURA AGIO Nasdaq argenxs lead oncology asset is ARGX-110 currently ”

It is important to note that the intention of the model is not to predict winning long term stocks, but to predict stock that will have a 10% increase within five business days.

BABA, a predicted “losers”, is also marked as a “buy” by all investing firms and reaffirms that additional data is needed for long term investments. This Snippet does show negative sentiment. Reducing holding in a company is not a good sign of positive things to come for a company. Even if this sentiment appears accurate, it does not on its own confirm the model’s accuracy. “ Tiger Legatus Capital Management Cut Alibaba Group Hldg LTD (BABA) Position By \$2.80 Million ... UtahHerald.com The company ”

## 4 CONCLUSION

The codes provided for this project take Google Alert data directly from a Gmail account, write the date the alert was received and the Snippet to a .csv, use the stock tickers identified in the Google Alerts to pull relevant historical stock price data to create a training set which is then analyzed using a random tree approach. The random tree analysis then produces a prediction for stocks that have received alerts more recently (within five days of the analysis). While all the sentiments drawn in the final calculation were indicated as “losers” none of the stocks were reconfirmed by recent historical data as significant increases. The lack of true negatives

**Table 4: Final analyzed results and accuracy**

Accuracy	Date	Sentiment	Ticker
L	2017-11-24	L	CLSN
L	2017-11-24	L	KMDA
L	2017-11-25	L	CLSN
L	2017-11-25	L	VTVT
L	2017-11-25	L	NKTR
L	2017-11-26	L	PRTD
L	2017-11-26	L	ADMA
L	2017-11-26	L	KALA
L	2017-11-26	L	SNDX
L	2017-11-26	L	GALE
L	2017-11-26	L	SNDX
L	2017-11-26	L	EVOX
L	2017-11-27	L	CPRX
L	2017-11-27	L	AZN
L	2017-11-27	L	TSRO
L	2017-11-27	L	CLSN
L	2017-11-28	L	MRK
L	2017-11-28	L	REGN
L	2017-11-28	L	EARS
L	2017-11-28	L	MRK
L	2017-11-28	L	CPRX
L	2017-11-28	L	AZN
L	2017-11-28	L	AERI
L	2017-11-29	L	CLSN
L	2017-11-29	L	EARS

does not confirm the model, but could be an indication of the model being on the right track for success.

The analysis presented herein represents the possible impact of sentiment expressed in news reports about clinical trials has the potential to predict the movement of stock prices. Further analysis should work with a bigger data set, possibly by increasing the number of configured Google Alerts and certainly by identifying how to pull stock tickers from the Snippets. An idea to do this might be to create a dictionary of stock tickers and company names and compare this dictionary with the sentiments. This could then pull out any company names or tickers defined in the Snippets and associate the relevant ticker symbol.

Next steps should also include more in depth analysis on the timing of stock increases by changing the historical stock data from five days after an alert is received to two days or one day. This would allow for a more immediate reflection on the cause and effect of the reported news. The scope should also be scaled to consider historical data over six months or more and compared again to the results of dedicated investor houses. In addition, adding sentiment analysis reports for pharmaceutical companies may benefit the long and short term predictions.

This project was run on ubuntu and took approximately four minutes to process from pulling Google Alerts to producing the analysis after Nltk was downloaded. Nltk took some seven minutes to download for the first run. Future projects, with bigger datasets, could be run from cloud environments like AWS, Chromeleon, or the server node of a big red environment.

**Table 5: Final analyzed results and accuracy continued**

Accuracy	Date	Sentiment	Ticker
L	2017-11-29	L	MRK
L	2017-11-29	L	TEVA
L	2017-11-29	L	NVS
L	2017-11-29	L	TEVA
L	2017-11-29	L	MRK
L	2017-11-29	L	EARS
L	2017-11-29	L	APVO
L	2017-11-29	L	EARS
L	2017-11-29	L	CLSN
L	2017-11-29	L	EARS
L	2017-11-30	L	ACAD
L	2017-11-30	L	MRK
L	2017-11-30	L	TEVA
L	2017-11-30	L	VKTX
L	2017-11-30	L	MRK
L	2017-11-30	L	TEVA
L	2017-11-30	L	VKTX
L	2017-11-30	L	CLSN
L	2017-11-30	L	NVS
L	2017-12-01	L	ABBV
L	2017-12-01	L	BAYN
L	2017-12-01	L	CLSN
L	2017-12-01	L	CLSN
L	2017-12-01	L	NTNX
L	2017-12-01	L	PAIOF

Continued improvement of the code would test running Kaggle and Nltk from the Google API script to reduce the size of the output file by eliminating stop words and special characters before the first export is even produced. This process would also improve speed with the historical stock price collection script as the Snippets are also written here.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the teaching assistants of the Fall 2017 i523 course for their support and suggestions in writing this paper.

## REFERENCES

- [1] Seeking Alpha. 2010. How Analyst Recommendations Affect Stock Prices: New Research. Website. (03 2010). <https://seekingalpha.com/article/194435-how-analyst-recommendations-affect-stock-prices-new-research>
- [2] G. Armano, M. Marchesi, and A. Murru. 2005. A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences* 170, 1 (2005), 3 – 33. <https://doi.org/10.1016/j.ins.2003.03.023> Computational Intelligence in Economics and Finance.
- [3] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1 – 8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- [4] F. Braudel. 1982. *Civilization and Capitalism, 15th-18th Century, Vol. II: The Wheels of Commerce*. University of California Press, California. <https://books.google.com/books?id=WPDbSXQsvGIC>
- [5] Adit Deshpande. 2017. Perform sentiment analysis with LSTMs, using TensorFlow. Website. (07 2017). <https://www.oreilly.com/learning/perform-sentiment-analysis-with-lstms-using-tensorflow>
- [6] London Stock Exchange. 2017. About London Stock Exchange Group. Website. (01 2017). <https://www.lseg.com/about-london-stock-exchange-group>

**Table 6: Final analyzed results and accuracy continued**

- | Accuracy | Date       | Sentiment | Ticker |
|----------|------------|-----------|--------|
| L        | 2017-12-01 | L         | SNDX   |
| L        | 2017-12-01 | L         | SNY    |
| L        | 2017-12-01 | L         | SNY    |
| L        | 2017-12-01 | L         | CLSN   |
| L        | 2017-12-01 | L         | NTNX   |
| L        | 2017-12-01 | L         | ABBV   |
| L        | 2017-12-02 | L         | ALNY   |
| L        | 2017-12-02 | L         | BABA   |
| L        | 2017-12-02 | L         | CISN   |
| L        | 2017-12-02 | L         | CLSN   |
| L        | 2017-12-02 | L         | MDXG   |
| L        | 2017-12-02 | L         | OCRX   |
| L        | 2017-12-02 | L         | PTCT   |
| L        | 2017-12-02 | L         | RTRX   |
| L        | 2017-12-02 | L         | CISN   |
| L        | 2017-12-02 | L         | RTRX   |
| L        | 2017-12-02 | L         | CLSN   |
| L        | 2017-12-03 | L         | ABBV   |
| L        | 2017-12-03 | L         | ARGX   |
| L        | 2017-12-03 | L         | BPMX   |
| L        | 2017-12-03 | L         | CLSN   |
| L        | 2017-12-03 | L         | CLSN   |
| L        | 2017-12-03 | L         | CSX    |
| L        | 2017-12-03 | L         | GERN   |
| L        | 2017-12-03 | L         | OMER   |
| L        | 2017-12-03 | L         | SGEN   |
| L        | 2017-12-03 | L         | SPHRF  |
| L        | 2017-12-03 | L         | TILE   |
| L        | 2017-12-03 | L         | TJX    |
- 
- [7] L.M. Friedman, C. Furberg, and D.L. DeMets. 1998. *Fundamentals of Clinical Trials*. Springer, Switzerland. <https://books.google.com/books?id=yzxT0Zh3X3IC>
  - [8] FXCM. 2014. New York Stock Exchange. Website. (12 2014). <https://www.fxcm.com/insights/new-york-stock-exchange-nyse/#history>
  - [9] O. Gassmann, G. Reepmeyer, and M. von Zedtwitz. 2013. *Leading Pharmaceutical Innovation: Trends and Drivers for Growth in the Pharmaceutical Industry*. Springer Berlin Heidelberg, Germany. <https://books.google.com/books?id=4Za-BwAAQBAJ>
  - [10] Google. 2017. Easily access Google APIs from Python. Website. (01 2017). <https://developers.google.com/api-client-library/python/>
  - [11] Google. 2017. Google Alerts. Website. (2017). <https://www.google.com/alerts>
  - [12] hugovk. 2017. Httplib2. Website. (10 2017). <https://github.com/httplib2/httplib2>
  - [13] Investopedia. 2017. Stock Market. Website. (09 2017). <https://www.investopedia.com/terms/s/stockmarket.asp?gl=rira-layout>
  - [14] Douglas B. Kitchen, Hlne Decornez, John R. Furr, and Jrgen Bajorath. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* 3 (Nov. 2004), 935. <http://dx.doi.org/10.1038/nrd1549>
  - [15] LINFO. 2006. the tar Command. website. (07 2006). <http://www.linfoo.org/tar.html>
  - [16] Steven Loria. 2017. TextBlob. Website. (01 2017). <http://textblob.readthedocs.io/en/dev/quickstart.html>
  - [17] Lukaszbanasiak. 2016. Yahoo-finance. Website. (12 2016). <https://github.com/lukaszbanasiak/yahoo-finance>
  - [18] J.P.Morgan Asset Management. 2017. Guide to the Markets. Website. (2017). <https://am.jpmorgan.com/us/en/asset-management/gim/adv/insights/guide-to-the-markets/viewer>
  - [19] NASDAQ. 2017. NASDAQ DataOnDemand Subscription Plans. Website. (2017). <https://www.nasdaqdod.com/Shop/ProductConfig.aspx?product=webservices&service=NASDAQDataOnDemand>
  - [20] NASDAQ. 2017. NASDAQ's Story. website. (2017). <http://business.nasdaq.com/discover/nasdaq-story/index.html>

```
bibtext report
```

```
=====
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "Helverbatirom"
(There was 1 warning)
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
[2017-12-16 09.36.35] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex
p.1 L33 : [Helverbatirom] undefined
Missing character: ""
Text page 6 contains only floats.
Text page 6 contains only floats.
Missing character: ""
There were undefined citations.
Typesetting of "report.tex" completed in 1.1s.
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Tiffany Fabianac
hid: 313
paper1: Oct 31 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
11
wc 313 project 11 7167 report.tex
wc 313 project 11 7275 report.pdf
wc 313 project 11 982 report.bib
```

```
find "
```

---

```
101: {"installed": {"client_id": "#",
102: "##.apps.googleusercontent.com",
103: "project_id": "#",
104: "auth_uri": "[URL]",
105: "token_uri": "[URL",
106: "auth_provider_x509_cert_url": "[URL",
107: "client_secret": "#",
108: "redirect_uris": ["urn:ietf:wg:oauth:2.0:oob",
109: "http://localhost"]}}
252: temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")
261: temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")
390: clean_train_reviews.append(" ".join(KaggleWord2VecUtility.
393: vectorizer = CountVectorizer(analyzer="word", tokenizer=None,
```

```
    preprocessor=None,  
  
403: #'Cleaning and parsing  
  
405: clean_test_reviews.append(" ".join(KaggleWord2VecUtility.  
  
410: print "Predicting test labels...\n"  
  
passed: False  
  
find footnote  
-----  
  
passed: True  
  
find input{format/i523}  
-----  
  
passed: False  
  
find input{format/final}  
-----  
  
6: \input{format/final}  
  
passed: True  
  
floats  
-----  
  
67: NASDAQ's website provides historical stock performance data that  
    can be exported as a Comma-Separated Values (CSV) file. The  
    disadvantage of NASDAQ's free export service is that each stock  
    must be exported separately. The free quote service can be  
    accessed at \cite{www-quotenasdaq}. NASDAQ provides API services  
    for subscribers starting at \$5,000 per year \cite{www-nasdaq-  
    sub}. Access to NASDAQ's API services can also be granted through  
    corporate sponsorship. NASDAQ's free CSV export services were used  
    to collect initial project data. In this example, the stock  
    history for Celsion Corporation during the week of August 21, 2017  
    is shown in \ref{T:nasdaq}.  
69: \begin{table}[htb]  
70: \caption{NASDAQ CSV file format example}\label{T:nasdaq}  
97: To collect data from the received Google Alerts without too much  
    manual clicking, Gmail has an available API which allows users to  
    pull data from a Gmail account. To start using the Gmail API, a
```

user must first configure their Authentication credentials through Google's developer console. The JSON format is shown in Table \ref{T:gapi}.

98: \begin{table}[htb]

99: \caption{Google Gmail API JSON format}\label{T:gapi}

181: \caption{The Google API Python code calls the Gmail APIs Messages.list which lists reduced properties of Gmail messages and Messages. Get which returns the messages themselves. Lists is used to query the messages that are wanted based on the defined criteria: userId=me, labelIds=INBOX], q=from:googlealerts-noreply@google.com. Get then retrieves the messages identified in using List and returns the messages content for Date and Snippet.}\label{c:googleapi}

184: Figure \ref{c:googleapi} shows the entire code to extract Google Alerts data using the Google provided Gmail API.

268: \caption{This Python script takes in the Date, Stock Ticker Symbol, and Snippet from the Google API .csv that was produced using both manual mining of the stock symbols and the python script provided for getting the Date and Snippet from Gmail. This code returns a modified .csv which lists an ‘‘L’’ for stocks that did not increase by 10\% in five days and a ‘‘W’’ for stocks that increased by at least 10\%. It also prints the stocks that increased by at least 10\% along with the highest price over 5 days, the starting price on the day that the Google Alert was received, and the percent change.}\label{c:stock}

271: Figure \ref{c:stock} shows a portion of the code to combine the data produced by the Google Alert mining and available historic stock price data.

273: Twelve out of over one hundred stock tickers returned by Google Alerts were flagged as ‘‘Winners’’ for increasing in price by 10\% within five days after the Google Alert was received and are shown in Table \ref{T:win}.

274: \begin{table}[htb]

275: \caption{Winning Stock Tickers}\label{T:win}

413: \caption{The Sentiment Python code takes the .csv exported by the historical stock script and parses the Snippets to train on the stock script and apply it to more recent stock quotes and Google Alerts}\label{c:sentiment}

416: Figure \ref{c:sentiment} shows the entire code to train on the dataset provided by the historical stock data and Google Alert sentiments.

449: \caption{This Python script takes in the Date and Stock Ticker Symbol from the sentiment .csv that was produced using the sentiment python script provided for performing a random forest analysis on the Google Alert results. This code returns a modified .csv which lists an ‘‘L’’ for stocks that did not

increase by 10\% from the time the Alert was received to the current date and a ‘‘W’’ for stocks that increased by at least 10\%. It also prints the stocks that increased by at least 10\% and were marked as ‘‘winners’’ by the sentiment script.\label{c:result}

452: Figure \ref{c:result} shows a portion of the code to combine the data produced by the random forest analysis and combine it with available historic stock price data.

456: The resulting CSV file contains the accuracy of the prediction, if the stock did not increase by atleast 10\%, the date that the Google Alert was received, the Sentiment that was calculated by the random forest algorithm, and the stock ticker. The results export to a .csv as shown in Tables \ref{T:results}, \ref{T:results2}, and \ref{T:results3}.

457: \begin{table}[htb]

458: \caption{Final analyzed results and accuracy}\label{T:results}

489: \begin{table}[htb]

490: \caption{Final analyzed results and accuracy continued}\label{T:results2}

521: \begin{table}[htb]

522: \caption{Final analyzed results and accuracy continued}\label{T:results3}

```

figures 0
tables 6
includegraphics 0
labels 10
refs 8
floats 6

```

```

False : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are refered to: (refs >= labels)

```

```

Label/ref check
passed: True

```

```

When using figures use columnwidth
[width=1.0\columnwidth]
do not cahnge the number to a smaller fraction

```

```
find textwidth
```

---

passed: True

below\_check

---

WARNING: algorithm and above may be used improperly

365: Random Forest algorithms create decision trees for each variable. Each tree represents the sequence of events or decisions that led to the outcome or result. With each branch or step through the decision tree a probability is calculated for the outcome and the collection of trees work are combined to create multiple ‘‘regression lines’’ that are used to predict an outcome when presented with new data that does not have an outcome. The model or collection of trees form what is called a random forest can then be used to predict sentiment or outcomes. For stock data or other time series datasets, it is essential to continuously re-train the model to perform at its best. As mentioned above it is possible for additional models and even the model itself to begin to influence the prediction model.

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--I didn't find a database entry for "Helverbatiimrom"  
(There was 1 warning)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

-----  
passed: True

ascii  
-----

non ascii found 8211  
=====

The following tests are optional  
=====

Tip: newlines can often be replaced just by an empty line

find newline  
-----

passed: True

cites should have a space before \cite{} but not before the {  
-----

find cite {  
-----

passed: True

# How Big Data Will Help Improve People's Health Worldwide

Paul Marks

Indiana University

Online Student

Shepherdsville, Kentucky 40165

pcmarks@iu.edu

## ABSTRACT

Aside from people changing their habits, big data analytics may hold the best possibility for the improvement of worldwide health. It will enable the ability to correctly diagnose patients more quickly, even when the patients may not be able to be physically seen by a provider. It will be used to create treatment plans specific to not only an illness, but to the patient's overall health condition and history, demographics, environment, and access to resources. While it may not solve the problem of everyone not having access to the best of care, it can help to make sure everyone can get the best care possible for them. This paper explores the ways in which big data is evolving in the field of healthcare to make these possibilities become realities and looks at some of the social concerns which could hold it back.

## KEYWORDS

i523, hid327, healthcare, patient treatment, genomics, diagnosis

## 1 INTRODUCTION

There have been many advances in big data analytics over the last several years. More and more data is able to be processed in a shorter amount of time. There are also many new sources of data. Data is not what big data is about though. It is about taking data and turning it into information that can be useful. The application of big data can vary, but very few may be more important than the ability to use data for the betterment of people's health across the globe. This is one way in which data science can make a substantial contribution to humanity.

Making this a reality is not, nor will be, a simple task. Health data itself requires the proper handling of the information as it is very sensitive. On one hand people have a right to privacy. On the other, if data is kept isolated, not combined with records from other people, then this limits the ability to gather insight and find breakthroughs. The key is to ensure privacy, but keep the integrity and relationships of the data in order to preserve privacy while gaining insight. The insight gained has endless possibilities.

One issue facing the medical profession today is a lack of trained professionals. The number of patients per healthcare worker around the world can vary from more than six per 1,000 people to less than one half per 1,000[43]. It is easy to see how this one fact greatly impacts the expected lifespan of people. But what if a patient could be examined, diagnosed, and have access to a treatment plan without a human doctor needed? It may sound futuristic, but the technology is being implemented today thanks in part to data analytics.

The impact of big data on healthcare doesn't stop there. The cost of treating 5 percent of the most chronic conditions can consume up

to 50 percent of the money spent on healthcare[42]. One reason for this is prevention, diagnosis, and treatment plans are not optimized. There is not one way to help patients avoid chronic conditions. It is based on many inputs depending on the person, their environment, and other factors. These same aspects impact the effectiveness of treatment plans as well. One size does not fit all. Through analytics many factors are being analyzed along with the results of prior plans to determine which methods would be the most effective. Avoiding a chronic condition not only saves money, but extends a patient's life and improves the quality of their life.

The ability to take many factors into account for a patient goes well beyond chronic conditions. Genomic technology is progressing which is allowing for a person's individual genome to be one of the inputs. Each person on earth has their own specific genome with billions of combinations, some of which directly impact their health and susceptibility to illnesses. Through big data analytics, this type of analysis may one day be commonplace like taking blood pressure and other vital statistics into account.

The discovery of new drugs and how they can be used to treat people is being sped up by the power of big data techniques. Drug research requires an immense amount of information to be correlated and processed. Big data is helping to speed this up and even helps speed up clinical trials by matching the right set of circumstances to provide viable results.

Progress does not always come without drawbacks, and big data analytics in healthcare is no exception.

## 2 HANDLING THE DATA

### 2.1 Security

Any use of healthcare data must take into account the ability to protect the data. Therefore a brief understanding of the task must be addressed. Healthcare information usually has two forms of protected information: Personally Identifiable Information (PII) and Protected Health Information (PHI). In order to be able to keep data with this type of information you must follow very strict rules on safeguarding it. The best known regulations are based on the Health Insurance Portability and Accountability Act (HIPAA) of 1996. Among the governmental standards to comply with HIPAA are the Security Control Assessment[18] and Defense Information Systems Agency's Security Technical Implementation Guides[1]. These types of requirements can be costly and require constant changes to remain secure.

Even with the ability to secure the data properly, any company wishing to obtain data must have an approved reason to get the information or the approval of the patients involved. Obtaining approval from each patient in a big data application is not practical.

Data is needed from too many people to obtain approval for each of them. A common way to handle this is through de-identification.

De-identification is the ability to alter the data in such a way that you cannot link health information to a person or identify individuals in the data. However, in order for the data to be useful for analysis it cannot be changed randomly so the links between certain data elements from record to record is lost. For instance, a diagnosis of a specific cancer in a patient must still be able to be linked to treatment data, x-rays, blood tests, etc. from that patient. In other words, de-identification has to be done in such a way that the data integrity remains in place, but the individual's identity is protected. This can become complicated because data elements such as age, sex, and geographical location are important.

Fortunately there are software solutions to assist in the de-identification of medical information. The software is broken into two categories: structured data and free-form text. De-identification of structured data is generally easier. The data has a known set of fields of which the ones which can identify a person and their health are known. These fields are added to the software and algorithms are run against them. The resultant data is useful for analysis, but the identity of any individual is safe. This is because the algorithm changes data in such a manner that it protects the person and the data integrity. Examples of tools in this arena include PARAT from Privacy Analytics, Inc., mu-Argus from the Netherlands national statistical agency, Cornell Anonymization Toolkit (CAT), an anonymization toolkit from the University of Texas at Dallas, sdcMirco from r-project.org[25]. Commercial tools like Privacy Analytics Eclipse claim to de-identify 10 million records per day from a variety of sources[50].

Unstructured data is more complex. The data which needs to be de-identified can be located anywhere within the dataset. This includes the text or metadata attached to images such as x-rays. Vital clinical, diagnosis, treatment, and other medical information is also included throughout unstructured data. Not being able to identify all PHI and PII can cause privacy concerns. Not linking all the correct data together reduces data integrity which reduces the usefulness of the data being studied.

Being able to properly de-identify and link unstructured data is being studied and refined. There are challenges for solutions to the problem. Informatics for Integrating Biology and the Bedside[24] has held challenges to help further solutions for this problem. The most recent was held in 2014. Track 1 of this challenge noted that "Removing protected health information (PHI) is a critical step in making medical records accessible to more people, yet it is a very difficult and nuanced"[24]. The ability to properly de-identify the data is rooted in the ability for the software to perform natural language processing. The focus of the challenge was all eighteen HIPAA defined PHI types[35]. While not as mainstream as de-identifying structured data, the ability to de-identify unstructured data will continue to progress and be solved through commercially available products over time.

## 2.2 Data Sharing

There are many sources of healthcare data. This is a major hurdle as the data is in different systems which are governed by different entities and used for purposes[32]. Data is stored in claims systems,

clinical settings, pharmacies, and others. It is stored in different formats. These sources may not contain similar key data that allows it to be easily brought together. Individual patients usually have a single provider who is their primary insurer. This data is usually in standard formats. However the same patients may have many providers of care using different systems. While most providers leverage electronic health records, these systems can contain many free-form text fields, images, and other types of fields. These data sets contain a wealth of information, but they are missing data which could be vital such as social, environmental, and community data. Other sources of data which could be useful are habits which people store on themselves such as food and activity tracking they may enter into any number of online applications[10].

While more data is being collected, there are still barriers to sharing it. There are the security and privacy concerns discussed earlier, but also the costs and who pays for them which must be addressed. There are tools and strategies being worked on in the industry to make sharing data across disparate systems possible. So far a widely adopted solution has not emerged[9]. Until such time that it does, data analytics in healthcare will be hampered.

## 3 BIG DATA IN A CLINICAL SETTING

Being a doctor can be like being a human big data machine at times. They take in many variables, process it against the history of information they have, and come to some sort of conclusion. In many cases there are multiple diagnosis that can be made. In fact sometimes there a lot of diagnosis that can be made. Unfortunately while much of the work is very scientific it does not mean that coming to a conclusion is a precise science.

Different doctors have different backgrounds. They have seen different patients, seen different diseases, studied at different locations, and read different literature. In short, their diagnosis is based off of their experiences. Unfortunately experiences are a form of bias. It is not that someone is doing this on purpose for the betterment or detriment of someone, but it is how our brains are wired. Physicians are not immune to this and it can affect the ability to treat all patients and conditions equally or appropriately[11]. When set up correctly and fine-tuned over time, data analytics can minimize biases.

### 3.1 Electronic Health Records

The ability to use big data in a clinical setting is growing out of the movement to storing records electronically. Historically these records were stored in paper format. The amount of data to use for big data analysis continues to rise as adoption of Electronic Health Records (EHRs) increases. Countries such as Norway and the Netherlands adopted EHRs more quickly than others and were at 98% adoption by 2012. The United Kingdom (97% in 2012), New Zealand (97%), and Australia (92%) were early adopters as well.[13] The United States is potentially a large source of EHR information, but has lagged other countries when looking at adoption rates. However, by the end of 2016 over 95% of hospitals and over 60% of United States based doctors have achieved meaningful use certification for EHRs from the Centers for Medicare and Medicaid Services.[40] As all countries continue to move toward storing

health records electronically then the body of information available for analysis will grow.

### 3.2 Big Data as a Physician Assistant

What if each doctor had the collective knowledge of others? That could make for better and more accurate diagnoses around the globe. A doctor in the United States would have the knowledge of thousands of years of alternative medicine which may only be taught in schools in the far east. Not only is it possible, but big data is making it happen today through technologies such as IBM's Watson.

### 3.3 IBM's Watson Health

One of the challenges facing doctors today is the ability to keep up with changes in healthcare. Even doctors who specialize in a field cannot keep up with the amount of information that is being published. One estimate is that 8,000 medical journal articles are published each day[59]. This makes medicine a good fit for big data. Watson Health, IBM's name for their cognitive supercomputer focused on healthcare, is able to ingest millions of pages of information in seconds. This information becomes part of the core information Watson has at its disposal as it assists clinicians by offering recommendations for them to consider. In this way Watson is not the final decision maker, but helps doctors be better at what they do[29].

While Watson is delegated to a physician's assistant currently, it may not always be so. In order to test how accurate it is, IBM tried it on 1,000 patients. In this test Watson and the attending physician agreed 99 percent of the time. In fact, in 30 percent of the cases Watson offered pathways which the physician had not considered. Armed with information like this IBM believes that computer cognitive thinking will be mainstream in the next ten years[59]. Because of advances in other technology areas have been progressing so quickly, it is hard to disagree with them. For instance, computers are now able to instantaneously make decisions that seemed unimaginable just a few years ago which as lead to the realization of autonomous driving vehicles. The question may not be the technology, but if people will accept a diagnosis from a computer program such as Watson.

Watson was also tested to see how examining a patient's entire genome would be more beneficial than simply running a panel which focuses on a limited number of areas most commonly known to be related to the cancer a patient may be experiencing. While the cost of and speed of sequencing a person's genome has been reduced, there is still a lot of work to using this data for a specific diagnosis and treatment plan. Both Watson and team from the New York Genome Center analyzed a patient's genome. Each of them was able to identify gene mutations which would have pointed to a clinical trial or drug which may have been a better match than the treatment the patient received. The difference being that it took the team of physicians approximately 160 hours to come to their conclusion. Watson provided its results in 10 minutes[58]. While not perfect, Watson adds another tool doctors can leverage which would allow them to better diagnose and treat patients.

How does Watson do it? It is actually very similar to how a human doctor works. The patient's symptoms and other information is made available to Watson. From there it deduces the relevant elements and leverages any background information it may have such as patient and family history, labs, x-rays, and other test results. It then accesses other sources of information it has accumulated over time: treatment guidelines, relevant articles and studies, and potentially information from other patients similar to this patient. Watson develops hypotheses and runs them through a process to test its hypotheses and provide a confidence score for each. Watson then provides its recommended treatment options with its confidence rating to the physician[19].

One advantage of Watson, or any such system, is that every time it is used that patient is getting all of its collective knowledge. Today when a patient see a physician they are diagnosed by that physician and maybe one or two other people generally from the same office. However as Watson gets *trained* by specialists in such fields as Oncology, every doctor who uses Watson's assistance becomes as or more knowledgeable than the collective group. This means that each doctor is providing top of the field care even if they are being seen nowhere near a facility that is considered as the best world[36]. A patient in a country not seen as having world-class healthcare can get diagnosed as if they were at the Sloan-Kettering Cancer Center. It also means that a patient who may be seeing a specialist in one area may be diagnosed with an ailment outside of their field. This can save time in receiving the appropriate diagnosis and subsequent treatment which gives patients the best chance for recovery.

There are obstacles to making Watson available worldwide and that is the ability to understand different languages. Watson knows English, Brazilian Portuguese, Japanese, and Spanish and is learning others. As an example, IBM, the Cleveland Clinic, and Mubadala are teaming up and are building a hospital in the Middle East. The Cleveland Clinic is already a user of Watson Health and is expected to leverage that in the new facility as many chronic conditions in the United States are present in the Middle East as well. To prepare for this, IBM is teaching Watson Arabic[62]. As Watson learns more languages it will be able to be leveraged in areas around the world which that language is spoken allowing for those populations to advance their healthcare knowledge.

Another advantage that Watson has over human physicians is that it never forgets. Even doctors who try to keep up with changes in healthcare, they will never be able to remember information as precisely as Watson. And Watson is also consistent. A single doctor may be mostly consistent, but different doctors will provide different diagnoses given the same input. Watson will not unless it is programmed differently or new knowledge is ingested which can create a more accurate diagnosis. It also does not have bad days, get tired, and is available 24x7x365. Watson's incremental costs, the cost of using it for one or one million patients, is low. IBM has spent billions on it and is continuing to invest, but those costs will be spread out as usage goes up thus making Watson cheaper over time[19].

### 3.4 Implementing Big Data Diagnostic Systems

Leveraging such technologies can be implemented in various ways. The easiest way is to look at them as another tool in a physicians' tool chest. Once fully implemented the inclusion of big data assisted technologies will be seamless. Clinical information is being collected digitally on an increasing basis. As vital signs, x-rays, diagnostic images, lab results, and even discussions with the patients are collected digitally they will become part of the patient's electronic health record and the overall collective knowledge base. Watson or other software could provide insight to the physician. It may be present a collection of diagnoses scored in likelihood based on the evidence collected so far[28]. It could provide recommendations for next steps or information which could lead to a more complete recommendation.

The idea behind such a system, Watson or any similar tool, is to make physicians better through more accurate diagnosis. It allows for the use of big data without removing the human aspect of medicine. This will help to begin to include the big data and computer health diagnosis to patients who would otherwise not be open to it. For many people their relationship with their doctor is personal. They discuss items with their doctor they do not discuss with anyone else. They may not trust a computer with their health[28]. A non-caring, non-breathing inanimate object cannot be trusted with something so human. In this implementation a doctor would still be there providing the personal interaction with the patient and thus providing them with the best care including the collective knowledge of the system.

### 3.5 Replacing Doctors for Routine Visits

Having a doctor meet with a patient initially may not always be required. The ability for big data to leverage healthcare data could lead to helping alleviate the shortage of doctors and nurses in the United States and around the world. Worldwide there is an estimated shortage of skilled health professionals of 17.4 million of which 2.6 million are doctors. The problem does not get much better over time as the estimate for 2030 is over 14 million[45]. It takes a lot of time and money for a student to achieve the level of knowledge to fill these positions. Unless the students are already in the pipeline then there is not a good response to the problem. People cannot switch careers and be a doctor or a nurse in twelve months or some short time-frame.

Adding new big data doctors is simple. It is mostly a hardware problem. Buy the right equipment, install the right software, train the staff, and Dr. Data can see patients. Leveraging automated machines to take vital signs will free up time for staff[14] similar to how checking out via automated tellers at the grocery store has reduced the number of cashiers and baggers needed. A physical office offering virtual doctor's visits could be staffed with people trained on the technology more than medical professionals. They would be there to help make sure that people are using the machines correctly and to wipe down equipment after a patient has used it. A nurse would be there in case certain patients are unable to use the equipment and their information must be taken manually. They could also be there to take blood samples which would be processed by automated machines and included in the patient's profile.

Automated diagnosis systems are in use today in a limited basis. In the United Kingdom the National Health Service has approved the use of Your.MD (an AI powered mobile app) for diagnosis. When people are comfortable using a technology like this it limits the number of more basic cases a doctor has to see and allows them to concentrate on more difficult tasks. Another tool, Ada, learns a user's history, provides an assessment, and adds an option to contact an actual doctor if needed. Babylon Health takes it one step further by adding follow-ups with users to see how they are doing and can even set up a video consultation with a live general practitioner if needed[14].

## 4 LIMITING EPIDEMICS

Incorporating big data analytics into the healthcare environment has the ability to limit the spread of disease by taking current circumstances outside of the immediate patient into account. In a linked system data from other local, regional, national, and global patients can be leveraged. Are there other patients presenting similar circumstances? Did the other patients provide more details or mention something slightly different? Taking this into account may help to diagnose a specific person and to identify an outbreak of something. Is a disease spreading? Did patients come from a similar location such as a building? By being able to correlate this information immediately there is the potential to stop an outbreak from spreading thus saving an untold number of patients from pain and suffering and saving healthcare dollars by not having to treat more patients. Epidemics have an economic impact at many levels including "the micro (individual and household), meso (establishment, village or city) and macro (national and international)"[46].

## 5 INSURANCE

The option of having fully automated doctors' visits could alter the insurance market as well. Health insurance is about numbers. Actuaries spend time estimating the health of the consumers they cover and many other factors to determine what premium rates to set[38]. Insurers make a profit by taking in more money than the costs to administrate the plans and the cost of paying for claims combined. To reduce the costs of claims they set predetermined prices for services rendered by hospitals, physicians, and sometimes pharmacy companies. The lower they can drive the cost of the claims they cover the less they charge or the more money they make. Charging less can result in making more money as well as more people may choose to purchase coverage from that insurer.

By creating an option for autonomous doctor's visits or tele-medicine an insurance company could save money. The more methods can be deployed which can reduce overall healthcare costs, the less people will pay. There are multiple ways in which this can be included to reduce health insurance premiums, a high cost item for most people in the United States and other countries. Insurers can work with healthcare providers who leverage this technology to create a reimbursement policy that is less for services such as tele-medicine[33]. They could also offer plans to potential customers which require basic treatments to take place with autonomous or tele-medicine options before they go to a doctor's office. This would offer an economic advantage to people which in turn can not only lower costs, but help to increase the adoption of new technologies.

Such a system is not for everyone or every condition. The idea is not to replace all doctor's visits, but to allow those who are comfortable to take advantage of lower cost coverage. It will encourage younger people to keep insurance if it is made more affordable. Currently the highest rate of not having insurance in the United States is when someone can no longer be covered as part of their parents plan, starting around the age of 25[4].

## 6 PORTABILITY

More importantly than lowering the cost of healthcare or making seeing a doctor more convenient is the ability to make exceptional healthcare available almost anywhere. Big data using an automated doctor can have an impact on under-served areas the like of which no one has ever seen. Today there are people who do not have access to healthcare of any kind. When they get sick they may not have a place to turn. In developed countries the number of patients per doctor is generally in the low hundreds. In poor, *third world countries* the number of patients per doctor is in the thousands or tens of thousands[26]. There are people who try to help, such as Doctors Without Borders, by making visits to these areas to provide some support but it does not reach a level anywhere near what people in some countries have available to them. If each doctor could multiply their impact with technology then the under-served would be helped more. As technology advances so people could be seen by experts without one being physically present then even more people could be seen.

## 7 PATIENT DATA COLLECTION

### 7.1 Actual Data vs. Circumstantial Insight

The more valid data which can be collected on patients the better big data will be able to help improve treatment for people around the world. The more accurate the data, the more accurate the analysis and results will be. Fortunately technology is helping in this area as well. Many people around the world have access to devices which monitor different aspects of our daily lives. Hundreds of millions of people around the world have purchased wearable devices, many of which can be used to monitor activity and inactivity[57]. By the end of next year it is expected that over one-third of people in the world will own a smart phone which can also track this type of activity[56]. While they are not seen as a medical device, they can help to track activity which is useful for diagnosis and treatment. They are another input into the data about a patient which can be used to more accurately gather information. Today doctors rely on a patient to answer questions about their level of activity. With such a devices they can get a more accurate picture.

These devices are useful for more than just activity levels. They also provide insight into areas of people's lives they are not really able to answer accurately such as how they sleep. Many people may sleep they sleep well or not so well, but in fact they are basing this more on how they feel than how much rest and how good of rest they get. Activity trackers are able to track sleep patterns as well. They actively monitor your inactivity. When used correctly a wearer pushes a button to indicate they are going to sleep and when they get up in the morning. The monitor is then able to track how long it takes for someone to get into a motionless/restful state. It continues to track them throughout the night recording if they

move around, get up, etc. Getting good sleep is a key element of maintaining overall health[54].

More advanced features of activity trackers include the ability to monitor vital signs like heart rates. They can be extremely important to a diagnosis providing input similar to a mini stress test. This is especially true if a person exercises, such as during jogging. The device can monitor how far a person is moving and their associated heart-rate. By gathering this information, the data can be fed into patient's profile when they visit a doctor (virtually or physically) instead of having to wait for a patient to get a test done and receive that feedback. Shortening the time to collect data and accurately analyze the patient can be the difference between life and death.

One aspect of activity trackers which must be noted is their accuracy and consistency. This is something big data can help with as well. Steps from person to person are not of consistent stride, tracker accuracy changes from device to device, heart rate monitors vary, and sleep are not be tracked similarly across all products and types of activities[55]. Big data can help normalize this input so that it can become a reliable input. Analysis has been done on different monitors to see how accurate they are. In order to bring them into health analysis more tests can be performed to get an accurate picture of how the devices correlate to the actual distances walked and level of sleep.

Activity trackers are only the beginning. *Wearable technology* is an expanding field which is enhancing the collection of passive data. Sensors are being built into clothing which track more accurately and include more types of data[20]. This includes information like breathing rate and muscle activity. They not only collect more types of data, but can wirelessly transmit the data via Bluetooth[31]. This means they can create a more accurate picture based on electronic data which can be used as an input. The more this type of technology becomes commonplace, the more data which can be fed into a patient's health record and the collection of health information.

### 7.2 Follow-Up Visits

All of these devices also have the ability to not only be used in diagnosis, but in the monitoring of treatment plans. Is the patient exercising as they say they are? Is a medicine or other corrective action helping them to lower their heart rate or get more restful sleep? It can also help to notify the patient or doctor when they are exceeding a prescribed level of respiration or heart rate. This can trigger an alert for a patient if they are at risk or even that they may need to seek treatment. These levels will not only be set based on standards, but patient specific information[3]. They can also take into account the environment the person is in. Are they in a hot location or one with high allergy levels which could negatively impact them? This is what separates the treatment plans of today with those of tomorrow. Use the technology to more accurately collect data on the patient, use it to create a diagnosis, monitor the patient using the technology, feed that data back into the patient's health record, and adjust as needed based on factual information.

Beyond the use of commercially available monitoring systems, there are devices which collect data similar to the information collected by a physician. Simple systems such as a blood pressure monitors are common. Many other pieces of equipment can be prescribed by a physician for home monitoring. These systems

not only collect information, but are able to digitally transmit the data so that it can be automatically analyzed with other sources of information. A patient will get feedback without having to visit a doctor[3]. This helps to close another gap in healthcare which affect many people: not following up with their doctor. Missing these visits can negatively impact the patient. By easing the ability to be monitored, automating the data collection, and instantly analyzing that data will lead to better overall prognosis.

Big data will also help to change people's habits. By using the data collected a picture of potential outcomes can be made for a patient to contemplate. Instead of generalities, patients will receive advice based on their medical history, other patients like them, treatment plans, and other inputs based on the variables specific to the patient's circumstances. It can show a patient how they impact their recovery based on what they are doing or not doing. For instance if they miss taking their medicines on time, do not lose weight, continue to smoke, or whatever other variables they are in control of and how it affects their specific recovery or health status. Showing them in advance may give them the motivation they need to follow the plan more closely. Throughout their treatment the model can be updated based on the patient's actual adherence to the plan. This provides another feedback loop for the patient to course correct their habits if they have not been following it as outlined[3].

Not only will big data help to diagnosis patients more accurately, but it will also allow for the customization of treatment plans at levels not available today. Instead of relying on more general treatment plans, patients will have their plans customized by their specific set of circumstances. Demographic information about the patient will be used to compare to historical plans and outcomes of patients most closely related to their characteristics. This includes not only the patients themselves, but the environments they live in. Pollution, weather, access to ongoing care, income (the patient may have to work whereas a long period of rest would be better) and other circumstances will be variables which may not be controllable by the patient, but can be used to help treat them. The plan will not necessarily be the best treatment course, not everyone has the access to the best care or the ability to abide by it, but will instead be the best plan for them and their circumstances. Each patient will be able to maximize their chances of recovering or otherwise leading the most normal life possible.

## 8 ACCESS TO HEALTHCARE

It is estimated that over 400 million people do not have access to basic healthcare around the world and others are forced into extreme poverty because of what they pay for healthcare[47]. Through tools referred to as telemedicine, these numbers can be lowered. Telemedicine itself is the ability for people to get evaluated, diagnosed, and treated while the physician is not located where they are. When combined with a mobile diagnostic unit a patient can get similar care to someone who is seen at a clinic[52]. As advances in automated solutions such as IBM Watson evolve, there could be a day when these remote services are performed in very remote areas where communication with a physician would be technically challenging.

## 9 COST SAVINGS

Another reason why big data will be helping with healthcare more and more in the future is the most basic of reasons: Economics. Regardless of the country or political system, there is always an economic element which must be addressed. No country, no system has an endless supply of any services or funds. Because of that ideas which make the most economic sense have a better chance to be adopted. The economics of automating healthcare with big data analytics will reach a tipping point as time progresses.

Simply put, healthcare is getting more and more expensive every year and computing resources become cheaper every year. Worldwide the per capita expense of healthcare has risen from \$661 to \$1,059 (numbers in United States Dollars or USD) in the last 10 years[21]. That is a 60.21% increase in one decade. The average per capita may seem low to some but that is due to it being worldwide number. Many countries spend almost nothing on healthcare per capita while others spend thousands. For instance, in 2004 Vietnam spent \$30 USD per capita and \$142 USD in 2014. This is a 373% increase, but in total dollars it is still a fraction of \$6,369 (2004) and \$9,403 spent in the United States[21].

In contrast to this the cost of computing power has decreased year over year. Computer power is not as straightforward to analyze, but cost trends are easily seen. One way is to compare the cost using a baseline year and showing other years as a percentage of the cost of the baseline. Using December of 1997 as a baseline (100) of cost for computers, the cost of computers and peripherals in January 2004 had dropped to 16.2. In other words, to get the same amount of computer power in 2004 you only had to spend 16.2 cents for every dollar spent in December of 1997. By January of 2014 it had dropped to 4.9. Comparing the 2004 and 2014 numbers, the same ones used above for healthcare spending, the cost of computing had been reduced by 69.75%[41].

A specific component when it comes to big data is the cost of storage. The decline in the cost of storage over time is staggering. In the early 1980's the cost of one gigabyte (GB) of storage was in the hundreds of thousands of dollars. Using early 2004 as our baseline the cost for one GB of storage had dropped to just under \$2.00. By 2014 the cost had declined further to between three and four cents per GB[30]. The speed at which the data can now be retrieved as compared to 2004 is like comparing the speed of light to the speed of sound. Today's storage units are that much faster.

Using this data one can see that as we are able to leverage big data solutions to provide better healthcare we can also begin to slow the incline of healthcare costs and then lower the cost of healthcare over time. Adding a new virtual doctor will not take years of schooling which can cost hundreds of thousands of dollars in some countries. It will be the cost of some piece of common technology and a licensing fee for the software. As with most everything technology based, increasing the volume decreases the cost. So as more and more virtual doctors are brought online the cost of each will decrease.

## 10 CHRONIC CONDITIONS

Chronic conditions are ones that "are preventable, and frequently manageable through early detection, improved diet, exercise, and treatment therapy"[61]. They are also very expensive to manage

and treat. Worldwide in 2010 the total cost of heart disease alone was \$863 billion dollars (USD) and is expected to be \$1.44 trillion by 2030. Between 2011 and 2031 the cost of the top five chronic diseases (cancer, diabetes, mental illness, heart disease, and respiratory disease) will cost \$47 trillion (USD) globally[27].

It is not only the economic impact of chronic diseases that make them a target for big data analysis. Chronic diseases reduce people's quality of life. This cannot be factored into simple terms such as money. Chronic diseases are the cause of 60 percent of deaths worldwide[44]. In a 2002 study it was estimated that 84 percent of deaths were due to chronic diseases in Europe and Central Asia[12]. Chronic disease is so prevalent and impactful to people's lives that it has been labeled as "the most expensive, fastest growing, and most intricate problem facing healthcare providers in every nation on earth[7]." With data like this it is easy to see why advances in chronic diseases is important. The question becomes how do fight them.

## 10.1 Prevention

The best way to fight chronic disease is to never have one in the first place. The best way to reduce the number of people who get a chronic condition is early intervention. Big data analytics can be used to help with population health management when it comes to chronic diseases. That is by identifying those who are at a high risk of getting one of these costly, harmful conditions[7]. The ability to leverage big data in prevention is a two part process. First risk factors which are modifiable must be identified and then interventions need to be created which will have an impact on changing the factors[5].

Modifiable is the key word in the first aspect of using big data. A key to fighting many chronic conditions is for people to stop behaviors such as smoking, to eat healthier, and to exercise more. However, if it was as easy as letting people know this then there would be a lot less chronic disease already. Big data can take many factors into account and help to create a more precise message for a people with specific risk elements. For instance instead of telling a patient to eat more nutritious foods, by leveraging elements of their specific health factors a doctor can recommend more precise information such as asking them to include a particular dietary nutrient[5]. Big data can also help with the timing of the message. In a survey patients wanted more information from analytics that would have warned them before they developed a chronic condition[8]. When someone is presented with more personalized information (they are on a path and about to reach a point of no return) vs. general (a healthy lifestyle may prevent you having issues years down the road) they are more compelled to heed that information and act upon it.

Newer technologies outside of a clinical setting are helping to add to the data available to analyze and care for patients. Combining data from a patient's activity monitor, fitness tracking website, or food logs into their plan helps to create a feedback cycle for the healthcare provider. Many applications track food by scanning the USB code from the package. Making it simple helps to get people to do things. The easier it is, the more likely they are to do it. Taking this data and combining it with clinical data such as blood labs and vital statistics can show a patient how they are directly

impacting their health in a positive or negative manner. It changes the conversation from more of a public service announcement general message to one unique to them.

A special sub-section of patients are very high-cost patients. In the United States there are roughly five percent of patients who account for almost 50 percent of healthcare spending[6]. Identifying these patients and creating intervention plans that work can have an enormous impact on their lives and the cost of healthcare overall. Patients with seemingly similar risk factors may have very different prognoses. Obvious factors such as age, weight, sex, and vital statistics may be the same. In order for big data to help identify the five percent more data is needed. Including mental health data, genetic information, socioeconomic, marital status, living conditions, and even cultural factors into the analysis will allow for better predictions and better ways to intervene which will lead to better outcomes[6].

## 10.2 Management

Even with the best of preventive measures there will still be too many people with chronic conditions for years and decades to come. Approximately 25 percent of people with chronic conditions have restrictions in what tasks they can perform for themselves, at work, or at school[23]. Because of this big data must also be leveraged to help manage those with chronic conditions. Managing it is not only based on cost, but helping them to live a better quality of life with less trips to the doctors and less admissions to a hospital. Data analytics can help to customize treatment plans to the circumstances of each patient. It can see patterns in patient's data and help to determine better follow up schedules. This could mean the difference between a visit with their doctor or a costly hospitalization[2].

Part of the solution for using big data to help tackle chronic conditions is leveraging new sources of information from technologies such as wearables. As mentioned earlier they allow for real-time data to be collected, combined with other sources of information including that of other patients, and provide better treatment plans for patients. Historically the medical profession had to rely on subjective input from patients when they came in for a visit. How often were they active, did they log information like their heart rate and blood pressure when they should have. With some wearables all this information and more is gathered in real-time and can trigger an alert to a care management professional[23]. This means that changes can be made when they are needed and the patient can get immediate attention, not days or weeks later.

Another issue with chronic care for providers is that patients may have multiple conditions. They may be overweight, have diabetes, and hypertension. This leads a patient to having multiple doctors each working on a specific condition, but no real coordination across the diseases. A treatment for one condition may have a negative impact on the patient because of treatment or drugs prescribed for another condition. And this situation is not unique as there are many patients suffering from the same conditions simultaneously. Big data analytics can bridge this gap. By combining data from multiple sources, patients, and treatments physicians can create a customized treatment plan for a patient to combat all three illnesses in the best manner without adverse interactions[60].

The result of this is that big data can help people see that treatments are tailored to them and are making a difference. Data analytics allows for patient-centric care, not disease-centric care. Patient managers would work with patients providing details on their plan, their results, and will be able to show patients how the care plan affects their quality of life. It can help to create a healthcare environment “where patients are not only engaged in time but see improved health results at affordable costs”[53].

## 11 GENOMICS (PERSONALIZED HEALTHCARE)

The field of Genomics is investigating how healthcare can be more personal. How diagnosis and treatment plans will be based on a specific person instead of how the factors or ailment is normally seen and treated in the general population. This is essential work because in the United States up to 47 percent of the cost of healthcare is spent on interventions that do not provide any value. While the actual percentage may vary in other countries, this is a worldwide problem[29]. Any easy way to understand the difference is over the counter medicine. Generally speaking the instructions on a bottle are broken down into children and adults. Following the directions adults will take the same amount of medicine regardless of their age, weight, or overall health.

Genomics aims to make medicine very specific to an individual by breaking down each person’s genome. This is only possible through big data as a single person’s genome produces a lot of data because it has up to 25,000 genes with three million base pairs. One human genome can produce up to 100 gigabytes of data[17]. And the information from one individual is not what is required for personalized health. It requires genomes from many individuals. The more data available, the better the analysis can be on similarities between people and how they may react to certain treatments. This multiplies 100 gigabytes by thousands, then millions, then hundreds of millions.

Through advances in technology such analysis is possible. In 2003 the first human genome was sequenced. It was only after 13 years and approximately \$3 billion dollars. By 2015 the same work can be done in a few hours at a cost of just over \$1,000[37]. This means that more and more people can have their genomes sequenced and used for analysis and personalized diagnosis and treatment of diseases. As more and more genomes are collected and analyzed treatment can be based on their personal genome and their family traits through family based analysis. This analysis lets doctors see how people may have inherited a propensity to be susceptible to certain diseases based on mutations in their genomes. In addition, through population based analysis environmental and cultural factors can be included. It is estimated that by 2025 over 100 million genomes could be sequenced[22]. Analyzing the details of the building blocks of so many individuals will be an a big data challenge which can have an enormous impact on healthcare.

## 12 DRUG DISCOVERY

Discovering new drugs which can help us live a better life is something like finding a needle in a haystack. Large libraries of molecules have to be examined “against millions of data points spanning chemical, biological, and clinical databases”[15]. This is done looking for

relationships between diseases and drugs to see if a particular drug could be used to treat the disease. While the process is not new, this work is the basis of many new drug discoveries, the ability of current big data techniques speeds up the process allowing for drugs to be discovered more quickly[15].

One of the reasons for it being so complicated goes back to the discussion of the human genome: each person is a unique individual. If you have seen a commercial or advertisement for a prescription drug there is always a list, sometimes a very long list, of possible side effects. These are adverse impacts which can range from minor annoyances to death. Part of the challenge of drug discovery is attempting to identify and quantify the impact a drug may have on people. To speed this process healthcare big data has developed solutions such as array-based technologies which are purpose built to combinatorial problems. This lets researchers find patterns in the data more quickly, speeding up the overall process[16].

Once a drug is thought to have a potential positive use it must go through a testing phase before it is approved for use. This can be long process which has successes and failures. Big data is being used for “the improvement of clinical trial designs (e.g., endpoints, inclusion/exclusion criteria, etc.)”[34]. This not only allows for potentially a quicker time to market, and thus the ability help people sooner, but a cost savings without paying for trials which do not produce viable results.

## 13 INCENTIVES FOR ADOPTION

In the end many of the advances will only be possible if people accept them. So how can this number be influenced? The most logical way to do so is to make the adoption of these advances financially beneficial. People are more willing to take a chance when they can see a hard benefit. Insurance premiums can help to drive this and provide an immediate benefit. Plans could be offered in which a person’s primary care is provided by a big data doctor. People would have to consent to having their information stored electronically and compared against the data sets. Visits to physical doctors including for second opinions would be limited. They could even have different reimbursement models similar to preventive tests. Most insurance today covers preventive services at 100 percent and are not subject to a deductible. Electronic visits could be treated similarly. They could be covered at 100 percent, or some number higher than regular doctors visits, and may or may not be subject to a deductible. Leveraging these types of incentives will help to promote the use of advanced analytics in the healthcare field. As usage grows so will the basis of data available to analyze and the ability to create better analysis models.

Another incentive for leveraging big data analytics by physicians is being led by the governments and private insurance. Instead of paying for services as they are performed, alternate payment models are being explored. For instance, in the United States the Centers for Medicaid and Medicare services is creating Alternate Payment Models to stimulate high-quality, cost-efficient care[39]. Physicians are able to earn more income and profit by achieving better outcomes. They will be willing to invest in computer analysis which will help them to diagnose and treat patients better. The financial incentive will drive change in providers’ habits which will benefit the healthcare big data analytics and patients.

## 14 DRAWBACKS

Leveraging big data innovations does not come without hurdles. One of the first is that people are generally slow or not open to change. The more personal the need for change, the less open they are. Organizations (hospitals, physician groups) are no different. Part of being an individual is making choices based of what information you can gather and leveraging your ability to make a determination. This is part of what makes each person unique. It is also how we learn. The more we become dependent on machines, the less we store in our own brains and we stop “building the networks in our brains to solve a whole host of problems.[51]” As those in the healthcare field rely more on technology to diagnosis and treat patients, the less human innovation may leveraged which can have a detrimental effect over time.

A major complication in big data analytics in any setting is the quality of data. The term emphasis garbage in, garbage out has probably been applied to computer systems since the beginning. There are techniques used to combat this, but when it comes to people’s health it is a bit more important. A portion of the healthcare data used as a base for analysis comes from existing diagnosis and treatment performed by humans. In looking at second opinions for patients, it was estimated that “10% to 62% of second opinions yield a major change in the diagnosis, treatment, or prognosis”[49]. Extrapolating this number to the base of information in big data for analysis means that a significant portion of the data would be different if a patient simply went to a different doctor.

Aside from the data itself, there is the potential for the algorithms behind big data analysis to be biased or having discrimination built into them. There has been a lot of talk about a lack of diversity in the technology world, especially with companies in Silicon Valley. This lack of diversity could become manifested into the analytics behind healthcare analytics. Different cultures and different races have some unique healthcare challenges. With a lack of diversification in key jobs the developers of healthcare systems could under-serve large portions of the world’s population due to a lack of understanding of how certain diseases affect their everyday lives. The United States Federal Trade Commission has asked companies in general to look at how representative their big data is and whether their models have built in biases[48]. The fact that healthcare around the world varies based economic factors makes it easy to understand how the data itself can be discriminatory. More wealthy people will be proportionally more represented than the poor thus skewing the data toward conditions afflicting the wealthy.

While big data will help to diagnose patients and create treatment plans, it does not come without its drawbacks. One of the biggest may be innovation. Part of being human is the ability to think of what has not already been done before. As algorithms and data analysis based on the historical variables begin to become more commonplace, there will be a reduction in the human factor of the medical profession. When faced with what can seem like a dire situation, the human mind can think of new options not previously discovered. Trying something which may not seem to have an impact on the surface, but something completely unrelated to any prior decision made can lead to new alternatives. What will a computer do with a patient when it does not see any hope? A human physician may opt to take a risk. It is a well-informed risk

with the patient knowing that there are no guarantees. It is easy to assume when an automated course of action without a substantial chance of a positive outcome is encountered that a physician would be able to intervene. This is true for a while, but as more and more of medicine is turned over to computer diagnosis and treatments the pool of capable physicians will shrink. With less people involved the less chance there is that the truly gifted individuals who make strides in the field will even decide to enter the field in the first place. In other words, these individuals may decide on a different career path and their discoveries would be left undiscovered.

## 15 CONCLUSIONS

Big data is an expanding science in many fields. The ability to digitize, collect, store, and analyze data has never been more than it is today. The type of information that can be used in data analysis is expanding every day as well. Images, videos, and sound are all part of the inputs into big data. Computers are now able to leverage natural language processing to make inputs that much easier to collect. As this field continues to grow, the ability to leverage it in improving healthcare around the world will grow as well.

We are on the edge of a shift in healthcare for the betterment of humankind. Advances will not be limited to one nation or one class of people. While healthcare may not be universal in its application, not every person will be able to access the same level of care, there will be benefits which can eventually help all people. A mobile unit which can be taken to almost any part of the planet will be able to have the knowledge better than most doctors practicing today. Doctors will have access to new drugs, diagnostic information, and treatment plans than they ever had before. They will be able to leverage new advances in medicine without having to read as many publications as they can. They will have a tool that reads and learns for them and provides that insight on case by case basis.

Through the use of data analysis of sources of data which did not exist a decade or so ago, we will be able to identify when a disease is starting to spread and react, thus limiting its impact. Because of technology people will be spared from suffering and they will never even know it. By understanding the human genome people who may be more susceptible certain diseases can be treated before they take hold. Babies will have their genome sequenced while they are still in their mother’s womb. This one aspect of the power of big data, the ability to process and understand a human genome, may be the single largest breakthrough in healthcare. It can provide insight into how each person individually reacts to the world around them and what science can do to make that interaction better. What science can do to help each person avoid potential chronic conditions which are not only financially costly, but that severely reduce their quality of life or end their life. Through advances in big data we will not only live longer, but live better.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support throughout this process. By offering an environment in which students were able to explore areas in big data which interested them, we were all able to further our knowledge individually and collectively. This project is similar to big data itself. It brought together various thoughts which could be considered data points

into the collection of the class. With access open to all, and potentially future classes, the collection of projects becomes a big data collection unto itself.

## REFERENCES

- [1] Defense Information Systems Agency. [n. d.]. Security Technical Implementation Guides (STIGs). Online. ([n. d.]). <https://iase.disa.mil/stigs/Pages/index.aspx>
- [2] Rick Altinger. 2017. Five Big Data Solutions to Manage Chronic Diseases. Online. (08 2017). <https://medcitynews.com/2017/08/five-big-data-solutions-manage-chronic-diseases/?rf=1>
- [3] Geoff Appelboom, Elvis Camacho, et al. 2014. Smart Wearable Body Sensors for Patient Self-Assessment and Monitoring. Online. (2014). <https://archpublichealth.biomedcentral.com/track/pdf/10.1186/2049-3258-72-28?site=http://archpublichealth.biomedcentral.com>
- [4] Jessica Barnett and Edward Berchick. 2017. Health Insurance Coverage in the United States: 2016. Online. (09 2017). <https://www.census.gov/content/dam/Census/library/publications/2017/demo/p60-260.pdf>
- [5] Meredith Barrett, Olivier Humbert, et al. 2013. Big Data and Disease Prevention: From Quantified Self to Quantified Communities. *Big Data* 1, 3 (09 2013), 168–175. <https://doi.org/10.1089/big.2013.0027>
- [6] David Bates, Suchi Saria, et al. 2014. Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients. *Health Affairs* 33, 7 (2014), 1123–1131. <https://doi.org/10.1377/hlthaff.2014.0041>
- [7] Jennifer Bresnick. 2015. How Healthcare Big Data Analytics Is Tackling Chronic Disease. Online. (06 2015). <https://healthitanalytics.com/news/how-healthcare-big-data-analytics-is-tackling-chronic-disease>
- [8] Jennifer Bresnick. 2016. How Big Data, EHRs, IoT Combine for Chronic Disease Management. Online. (02 2016). <https://healthitanalytics.com/news/how-big-data-ehrs-iot-combine-for-chronic-disease-management>
- [9] Jennifer Bresnick. 2017. Top 10 Challenges of Big Data Analytics in Healthcare. Online. (06 2017). <https://healthitanalytics.com/news/top-10-challenges-of-big-data-analytics-in-healthcare>
- [10] Jennifer Bresnick. 2017. Which Healthcare Data is Important for Population Health Management? Online. (06 2017). <https://healthitanalytics.com/news/which-healthcare-data-is-important-for-population-health-management>
- [11] Elizabeth Chapman, Anna Kaatz, and Molly Carnes. 2013. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine* 28, 11 (11 2013), 1504–1510. <https://doi.org/10.1007/s11606-013-2441-1>
- [12] D'Vera Cohn. 2007. The Growing Global Chronic Disease Epidemic. Online. (05 2007). <http://www.pbs.org/Publications/Articles/2007/GrowingGlobalChronicDiseaseEpidemic.aspx>
- [13] ASC Communications. 2013. Top 10 Countries for EHR Adoption. Online. (06 2013). <https://www.beckershospitalreview.com/healthcare-information-technology/top-10-countries-for-ehr-adoption.html>
- [14] Ben Dickson. 2017. How Artificial Intelligence is Revolutionizing Healthcare. Online. (2017). <https://thenextweb.com/artificial-intelligence/2017/04/13/artificial-intelligence-revolutionizing-healthcare/>
- [15] Brian Eastwood. 2016. Bringing Big Data to Drug Discovery. Online. (09 2016). <http://mitsloan.mit.edu/newsroom/articles/bringing-big-data-to-drug-discovery/>
- [16] Suzanne Elvidge. [n. d.]. Digging for Big Data Gold: Data Mining as a Route to Drug Development Success. Online. ([n. d.]). <https://www.clinicalleader.com/doc/digging-for-big-data-gold-data-mining-as-a-route-to-drug-development-success-0001>
- [17] Bonnie Feldman. 2013. Genomics and the Role of Big Data in Personalizing the Healthcare Experience. Online. (08 2013). <https://www.oreilly.com/ideas/genomics-and-the-role-of-big-data-in-personalizing-the-healthcare-experience>
- [18] Centers for Medicare and Medicaid Services. [n. d.]. CMS Information Security and Privacy Overview. Online. ([n. d.]). <https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity/index.html?redirect=/InformationSecurity/>
- [19] Lauren Friedman. 2014. IBM's Watson Supercomputer May Soon be the Best Doctor in the World. Online. (04 2014). <http://www.businessinsider.com/ibms-watson-may-soon-be-the-best-doctor-in-the-world-2014-4>
- [20] Malarie Gokey. 2016. Why smart clothes, not watches, are the future of wearables. Online. (01 2016). <https://www.digitaltrends.com/wearables/smart-clothing-is-the-future-of-wearables/>
- [21] World Bank Group. [n. d.]. Health Expenditure per Capita (current US\$). Online. ([n. d.]). [https://data.worldbank.org/indicator/SH.XPD.PCAP?end=2014&name\\_desc=true&start=2004&view=chart](https://data.worldbank.org/indicator/SH.XPD.PCAP?end=2014&name_desc=true&start=2004&view=chart)
- [22] Karen He, Dongliang Ge, and Max He. 2017. Big Data Analytics for Genomic Medicine. *International Journal of Molecular Sciences* 18, 2 (02 2017), 18. <https://doi.org/10.3390/ijms18020412>
- [23] Scalable Health. 2017. Managing Chronic Conditions using Big Data. Online. (03 2017). [https://www.scalablehealth.com/Resources/WP/SS\\_Chronic\\_Illness-ThoughtPaper.pdf](https://www.scalablehealth.com/Resources/WP/SS_Chronic_Illness-ThoughtPaper.pdf)
- [24] Partners Healthcare. 2014. 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data. Online. (2014). <https://www.i2b2.org/NLP/HeartDisease/>
- [25] CHEO Research Institute. [n. d.]. What De-Identification Software Tools are There? Online. ([n. d.]). <http://www.ehealthinformation.ca/faq/de-identification-software-tools/>
- [26] Frank Jacobs. [n. d.]. The Patients Per Doctor Map of the World. Online. ([n. d.]). <http://bigthink.com/strange-maps/185-the-patients-per-doctor-map-of-the-world>
- [27] Kate Kelland. 2011. Chronic Disease to Cost \$47 Trillion by 2030: WEF. Online. (09 2011). <https://www.reuters.com/article/us-disease-chronic-costs/chronic-disease-to-cost-47-trillion-by-2030-wef-idUSTRE78H2IY20110918>
- [28] Bijan Khosravi. 2016. Will You Trust AI To Be Your New Doctor? Online. (03 2016). <https://www.forbes.com/sites/bijankhosravi/2016/03/24/will-you-trust-ai-to-be-your-new-doctor-a-five-year-outcome/#3629545b3724>
- [29] MS Kohn, J Sun, et al. 2014. IBM's Health Analytics and Clinical Decision Support. *Yearbook of Medical Informatics* 9, 1 (2014), 154–162. <https://doi.org/10.15265/IY-2014-0002>
- [30] Matthew Komorowski. 2014. A History of Storage Cost. Online. (03 2014). <http://www.mkomo.com/cost-per-gigabyte-update>
- [31] Max Langridge and Luke Edwards. 2017. Best Smart Clothes: Wearables to Improve Your Life. Online. (10 2017). <http://www.pocket-lint.com/news/131980-best-smart-clothes-wearables-to-improve-your-life>
- [32] Mona Lebied. 2017. 9 Examples of Big Data Analytics in Healthcare That Can Save People. Online. (05 2017). <https://www.datapine.com/blog/big-data-examples-in-healthcare/>
- [33] KJ Lee. 2017. Here's How to Reduce Healthcare Costs. Online. (05 2017). <http://medicaledconomics.modernmedicine.com/medical-economics/news/heres-how-reduce-healthcare-costs?page=0,1>
- [34] Lada Leyens, Matthias Reumann, et al. 2017. Use of Big Data for Drug Development and for Public and Personal Health and Care. *Genetic Epidemiology* 41, 1 (01 2017), 51–60. <https://doi.org/10.1002/gepi.22202>
- [35] Zengjian Liu, Yangxin Chen, et al. 2015. Automatic De-Identification of Electronic Medical Records using Token-Level and Character-Level Conditional Random Fields. *Journal of Biomedical Informatics* 58 (12 2015), S47–S52. <https://doi.org/10.1016/j.jbi.2015.06.009>
- [36] Laura Lorenzetti. 2016. Here's How IBM Watson Health is Transforming the Health Care Industry. Online. (04 2016). <http://fortune.com/ibm-watson-health-business-strategy/>
- [37] Sid Nair. 2015. How Advanced Genomics, Big Data will Enable Precision Medicine. Online. (09 2015). <https://healthitanalytics.com/news/how-advanced-genomics-big-data-will-enable-precision-medicine>
- [38] American Academy of Actuaries. 2016. Drivers of 2017 Health Insurance Premium Changes. Online. (05 2016). <https://www.actuary.org/content/drivers-2017-health-insurance-premium-changes-0>
- [39] Department of Health and Human Services. [n. d.]. APMs Overview. Online. ([n. d.]). <https://qpp.cms.gov/apms/overview>
- [40] United States Department of Health and Human Services. 2017. Health IT Dashboard. Online. (08 2017). <https://dashboard.healthit.gov/quickstats/quickstats.php>
- [41] United States Department of Labor. [n. d.]. Long-Term Price Trends for Computers, TVs, and Related Items. Online. ([n. d.]). <https://www.bls.gov/opub/ted/2015/long-term-price-trends-for-computers-tvs-and-related-items.htm>
- [42] Optum. [n. d.]. Data Rich, Insight Poor. Online. ([n. d.]). [https://cdn-aem.optum.com/content/dam/optum3/optum/en/images/infographics/Game\\_changer\\_Track.Two.04\\_Data\\_Rich\\_Insight\\_Poor\\_Infog\\_Images.2016.pdf](https://cdn-aem.optum.com/content/dam/optum3/optum/en/images/infographics/Game_changer_Track.Two.04_Data_Rich_Insight_Poor_Infog_Images.2016.pdf)
- [43] World Health Organization. [n. d.]. Density of Physicians (Total Number per 1000 Population): Latest Available Year. Online. ([n. d.]). [http://www.who.int/gho/health\\_workforce/physicians\\_density/en/](http://www.who.int/gho/health_workforce/physicians_density/en/)
- [44] World Health Organization. [n. d.]. Chronic Diseases and Health Promotion. Online. ([n. d.]). <http://www.who.int/chp/en/>
- [45] World Health Organization. [n. d.]. Global Health Observatory (GHO) data. Online. ([n. d.]). [http://www.who.int/gho/health\\_workforce/en/](http://www.who.int/gho/health_workforce/en/)
- [46] World Health Organization. 2005. Evaluating the Costs and Benefits of National Surveillance and Response Systems. Online. (2005). [http://www.who.int/csr/resources/publications/surveillance/WHO\\_CDS\\_EPR\\_LYO\\_2005\\_25.pdf](http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_LYO_2005_25.pdf)
- [47] World Health Organization and World Bank. 2015. New Report Shows that 400 Million do not have Access to Essential Health Services. Online. (06 2015). <http://www.who.int/mediacentre/news/releases/2015/uhc-report/en/>
- [48] Out-Law.com. 2016. Use of Big Data Can Lead to 'harmful exclusion, discrimination' fi!! FTC. Online. (01 2016). [https://www.theregister.co.uk/2016/01/08/use\\_of\\_big\\_data\\_can\\_lead\\_to\\_harmful\\_exclusion\\_or\\_discrimination\\_us\\_regulator/](https://www.theregister.co.uk/2016/01/08/use_of_big_data_can_lead_to_harmful_exclusion_or_discrimination_us_regulator/)
- [49] Velma Payne, Hardeep Singh, et al. 2014. Patient-Initiated Second Opinions: Systematic Review of Characteristics and Impact on Diagnosis, Treatment, and Satisfaction. *Mayo Clinic Proceedings* 89, 5 (05 2014), 687–696. <https://doi.org/10.1016/j.mayocp.2014.01.014>

- 1016/j.mayocp.2014.02.015
- [50] Inc. Privacy Analytics. [n. d.]. Privacy Analytics Eclipse. Online. ([n. d.]). <https://privacy-analytics.com/software/privacy-analytics-eclipse/>
  - [51] John Robison. 2009. Is Technology Making us Dumber? Online. (11 2009). <https://www.psychologytoday.com/blog/my-life-aspergers/200911/is-technology-making-us-dumber>
  - [52] Sameer Sawarkar. 2013. Remote Healthcare Solution. Online. (2013). <http://www.who.int/ehealth/resources/compendium.ehealth2013.7.pdf>
  - [53] Abhinav Shashank. 2016. Chronic Care Management Marries Big Data. Online. (12 2016). <http://blog.innovaccer.com/chronic-care-management-marries-big-data/>
  - [54] Alyssa Sparacino. 2013. 11 Surprising Health Benefits of Sleep. Online. (07 2013). <http://www.health.com/health/gallery/0,,20459221,00.html#go-ahead-snooze--1>
  - [55] Caitlin Stackpool, John Porcari, et al. 2015. ACE-sponsored Research: Are Activity Trackers Accurate? Online. (01 2015). <https://www.acefitness.org/education-and-resources/professional/prosource/january-2015/5216/ace-sponsored-research-are-activity-trackers-accurate>
  - [56] Statista. [n. d.]. Smartphones industry: Statistics & Facts. Online. ([n. d.]). <https://www.statista.com/topics/840/smartphones/>
  - [57] Statista. [n. d.]. Statistics & Facts on Wearable Technology. Online. ([n. d.]). <https://www.statista.com/topics/1556/wearable-technology/>
  - [58] Eliza Strickland. 2017. IBM Watson Makes a Treatment Plan for Brain-Cancer Patient in 10 Minutes; Doctors Take 160 Hours. Online. (08 2017). <https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ibm-watson-makes-treatment-plan-for-brain-cancer-patient-in-10-minutes-doctors-take-160-hours>
  - [59] Tom Sullivan. 2017. Cognitive Computing will Democratize Medicine, IBM Watson Officials Say. Online. (04 2017). <http://www.healthcareitnews.com/news/cognitive-computing-will-democratize-medicine-ibm-watson-officials-say>
  - [60] Ann Tinker. 2017. How to Improve Patient Outcomes for Chronic Diseases and Comorbidities. Online. (2017). <https://www.healthcatalyst.com/how-to-improve-chronic-diseases-comorbidities>
  - [61] Partnership to Fight Chronic Disease. [n. d.]. The Growing Crisis of Chronic Disease in the United States. Online. ([n. d.]). [https://www.fightchronicdisease.org/sites/default/files/docs/GrowingCrisisofChronicDiseaseintheUSfactsheet\\_81009.pdf](https://www.fightchronicdisease.org/sites/default/files/docs/GrowingCrisisofChronicDiseaseintheUSfactsheet_81009.pdf)
  - [62] Jonathan Vanian. 2015. IBM's Watson Supercomputer is Learning Arabic in Move to Middle East. Online. (07 2015). <http://fortune.com/2015/07/14/ibm-watson-home-middle-east/>

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty year in DISA
Warning--empty year in ClinicalLeader
Warning--empty year in CMS
Warning--empty year in WoldBankPerCapita
Warning--empty year in eHealthInfo
Warning--empty year in BigThink
Warning--empty year in CMSAPM
Warning--empty year in CompPrices
Warning--empty year in Optum
Warning--empty year in WHODensity
Warning--empty year in WHOChronicDisease
Warning--empty year in WHOGHO
Warning--empty year in PrivacyAnalytics
Warning--empty year in StatistaPhones
Warning--empty year in StatistaWearable
Warning--empty year in FightChronicDisease
(There were 16 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.37.45] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Missing character: ""
```

```
Typesetting of "report.tex" completed in 1.1s.
```

```
=====
```

## Compliance Report

```
=====
```

```
name: Marks, Paul
hid: 327
paper1: 100% 10/25/2017
paper2: 100% 11/06/17
project: 100% 12/05/17
```

```
yamlcheck
```

---

```
wordcount
```

---

```
11
wc 327 project 11 10029 report.tex
wc 327 project 11 10764 report.pdf
wc 327 project 11 2080 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

floats

---

figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0

True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
True : check if all figures are refered to: (refs >= labels)

Label/ref check  
passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst

```
Database file #1: report.bib
Warning--empty year in DISA
Warning--empty year in ClinicalLeader
Warning--empty year in CMS
Warning--empty year in WoldBankPerCapita
Warning--empty year in eHealthInfo
Warning--empty year in BigThink
Warning--empty year in CMSAPM
Warning--empty year in CompPrices
Warning--empty year in Optum
Warning--empty year in WHODensity
Warning--empty year in WHOChronicDisease
Warning--empty year in WHOGHO
Warning--empty year in PrivacyAnalytics
Warning--empty year in StatistaPhones
Warning--empty year in StatistaWearable
Warning--empty year in FightChronicDisease
(There were 16 warnings)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Big Data Applications in Predicting Hospital Readmissions

Tyler Peterson

Indiana University - School of Informatics, Computing, and Engineering

711 N. Park Avenue

Bloomington, Indiana 47408

typeter@iu.edu

## ABSTRACT

Hospital readmissions occur when a patient is discharged from a hospital and subsequently readmitted to a hospital within a short time frame. Hospitals are held accountable and penalized for readmissions that occur within 30 days of the initial inpatient stay. In 2016, nearly 2,600 hospitals were penalized \$528 million collectively for readmissions. Machine learning is increasingly being used to build models that predict if a patient has a high probability of being readmitted, which allows hospital staff to prioritize resources around high-risk patients and potentially prevent the otherwise likely readmission. Healthcare providers possess every-growing stores of medical data that are essential for building accurate predictive models. While most of this information is private and not widely available for research, there are a few public datasets that researchers can use to build models and gain a better understand of which information is significant in the task of identifying high-risk patients. One such dataset includes over 100,000 patient admissions that occurred at 130 US hospitals between 1999 and 2008 and includes many features that can be used to build models. Open-source Python tools such as scikit-learn, pandas and matplotlib have tools necessary for preparing, modeling and visualizing data. These tools can be used to define algorithms that describe the problem of hospital readmissions by creating classifiers that categorize patients based on the probability of readmission. Machine learning techniques, such as logistic regression, are capable of modeling data for classification problems, and these tools include methods for assessing and optimizing the algorithms. In this analysis, the model created using logistic regression performed better than random guessing, but not well enough to reasonably be considered a highly effective model. The sensitivity of the model is rather low for a problem where there is a high cost of missing an opportunity to intervene on a patient at high-risk of readmission. The lack of behavioral and social attributes in the dataset may lend to lower predictive power. In any case, the effectiveness of machine learning in classifying patients for risk of readmission is a growing topic of study and implementation of tools for assisting healthcare providers will likely continue to increase.

## KEYWORDS

hid331, i523, Big Data, Hospital Readmissions, Machine Learning, Classification, Python

## 1 INTRODUCTION

Hospital readmissions are problematic for both patients and health-care providers. Even a single hospital admission for a patient can be an inconvenient, expensive and anxiety-inducing major life event.

For a patient to be subsequently readmitted to the hospital, the patient again experiences the negative aspects of being in a hospital, along with a diminished quality of life that accompanies a recurrent disease or medical issue. Healthcare providers are increasingly being held accountable and often penalized for an inability to keep recently discharged patients from being readmitted. It has been estimated that nearly 1 in 5 Medicare patients discharged from a hospital will be readmitted within 30 days [5].

The Hospital Readmission Reduction Program (HRRP), which originated in 2013 as a provision in the Affordable Care Act, serves as an example of an initiative that punishes hospitals for readmissions by administering financial penalties on hospitals with disproportionately high readmission rates among Medicare beneficiaries [1]. The HRRP levies a reduction in Medicare reimbursement, and uses the ‘all-cause’ definition for readmissions, which means that a subsequent hospital stay that occurs for any reason within 30 days of the initial stay counts against the hospital [1]. The program focuses on patients initially admitted with a heart attack, heart failure, pneumonia, chronic obstructive pulmonary disease, a coronary artery bypass graft procedure or a hip/knee replacement procedure [1]. If a hospital’s risk-adjusted readmission rate is higher than the national average, then that hospital will be penalized. Further, the excessiveness of the rate is considered as well, ensuring that providers with the worst readmission rates have proportionately higher penalties [1]. In 2016, the US government penalized 79 percent of US hospitals, which amounts to 2,597 institutions [9]. The penalties for those readmissions, applied to the 2017 fiscal year reimbursements, amounted to \$528 million nationally, \$108 million higher than the previous year [9].

Effectively this means that the care provided to readmitted patients is uncompensated care, which still requires valuable resources such as medical supplies, pharmaceuticals, the occupancy of hospital beds and the attention of medical staff. HRRP has had the intended effect of bringing increased attention to readmissions, and some healthcare providers are leveraging their ever-increasing medical data stores to better understand their patients. Several organization are using machine learning to identify high-risk patients. Assessing patients for the likelihood of readmission presents a binary classification problem, where a model’s goal is to come to one of two conclusions on each case. The model analyzes each patient and the patient’s accompanying attributes and concludes either that the patient will be readmitted or will not be readmitted.

### 1.1 Applying Machine Learning to Hospital Readmissions

There are several studies pertaining to the effectiveness of using machine learning to build predictive models that address this problem.

A 2011 study conducted a systematic review of the topic and found 26 studies discussing predictive models related to hospital readmissions. These models were created using administrative claims data, electronic medical record (EMR) data, or a combination of each type of dataset [4]. Administrative claims data is primarily gathered for billing purposes and contains information about procedures, diagnoses, length of hospital stay and location of care [7]. The advantage of this type of data is that it typically describes large populations and is inexpensive to acquire because it's already gathered for billing [5]. EMRs contain the basic information contained in administrative claims data, and also include lab data, image data and the results of various diagnostic tests, as well as social and behavioral information. Of the 26 studies reviewed by this paper, only 4 reported an area under the curve (AUC) value greater than 0.70, indicating that the other 22 models performed relatively poorly at classifying high-risk patients. Interestingly, 3 of the 4 studies with a moderately high AUC built models with clinical information found in EMRs in addition to administrative claims data, which suggests that the rich information available in EMRs adds discriminative power to the predictive models [5].

One study that demonstrates the power of incorporating EMR data was conducted at Mount Sinai Health System in New York, NY. Mount Sinai developed a model to predict readmissions among patients with heart failure, which is the top cause of readmission among Medicare beneficiaries [10]. To build the model, Mount Sinai leveraged their EMR system to mine 4,205 patient attributes, including 1,763 diagnosis codes, 1,028 medications, 846 laboratory measurements, 564 surgical procedures, and 4 types of vital signs. The study used a cohort of 1,068 patients, 178 of whom were readmitted within 30 days [10]. The model achieved a prediction accuracy rate of 83.19 percent and an AUC value of 0.78. Commenting on this outcome, Mount Sinai said that the model would benefit from the inclusion of several years of data from several different hospital sites [10]. In other words, even more data is needed to further improve the accuracy of the model.

## 2 ANALYSIS

Though the data used by institutions to build models is not widely available, there are a few public datasets that can be used by machine learning practitioners to better understand how predictive modeling techniques can be applied to the task of predicting readmissions. One such dataset comes from the Cerner Corporation's Health Facts database, which is comprised of comprehensive clinical EMR records voluntarily provided by hospitals across the United States [11].

Researchers extracted a subset of 101,766 encounters from the nearly 74 million records in the Health Facts database for the purpose of studying diabetic inpatient encounters. The admissions span 10 years from 1999 to 2008, and occurred at 130 different hospitals across the United States. The researchers used the following criteria to narrow down the dataset [11]:

- 1) The encounter is an inpatient encounter.
- 2) It was a diabetic encounter, meaning at least one diabetic diagnosis code was associated with the episode of care.
- 3) The length of stay was between 1 and 14 days.
- 4) The patient had at least one lab test.

- 5) The patient was administered at least one medication.

This dataset is now publicly available on the UCI Machine Learning Repository. Each observation in the dataset has up to 55 attributes, or features, that are potentially related to hospital readmissions, including diagnoses defined by ICD9 codes, in-hospital procedures, hospital characteristics, individual provider information, lab data, pharmacy data, and demographic data, such as age, gender and race. Each patient encounter record also has a label indicating whether or not the patient was readmitted within 30 days. Since the dataset includes these labels, supervised machine learning techniques can be used, as opposed to unsupervised machine learning techniques. Logistic regression is a supervised machine learning technique capable of binary classification of observations, and is well-suited to predict the likelihood of readmission for the observations in this diabetes dataset.

### 2.1 Overview of Supervised Machine Learning

**2.1.1 Minimization of Error.** The goal of a machine learning algorithm is to minimize the error made in the predictions. The general form of this concept can be represented by the formula:

$$Y = f(x) + \epsilon$$

$Y$  is the actual outcome associated with the sample.  $x$  represents the attributes associated with each sample and typically takes the form of a matrix where the columns are the features and the rows are the individual observations.  $f(x)$  is a function that represents the systematic information  $x$  provides about  $Y$ , and  $\epsilon$  is the error term describing the differences between the predicted value returned by  $f(x)$  and the actual value represented by  $Y$  [6]. A perfect prediction means  $f(x)$  equals  $Y$  and  $\epsilon$  equals zero. In reality, the error term will rarely be zero, so each prediction yields a certain amount of error. The prediction accuracy for each sample is evaluated by this formula, and sum of the error terms from each evaluation represents the magnitude of error made by the model. The goal is to make the sum of errors as low as possible [6].

The error term is minimized through optimization of  $f(x)$ , which is intended to describe the patterns that exist between the independent variables, represented by  $x$ , and the dependent variable, represented by  $Y$ . Said differently, the equation describes the relationship between the features and the outcome label. The way that this function describes this relationship is through coefficient weights. Each feature in the dataset is paired with a numerical weight that accentuates or diminishes the impact of a feature on the predicted outcome. The way in which these coefficients can be interpreted differs by which algorithm is used, but the intuition remains the same: the coefficients are adjusted to highlight the important features in the dataset. Once the coefficients are determined, the model has been fit to the data.

**2.1.2 Training Set vs. Test Set.** The coefficient weights of the model are defined by analyzing the samples in a dataset. In a practical sense, the value of a model depends on its ability to accurately predict the outcomes of new samples that were unseen at the time the model was determined [4]. A model that performs well when making predictions with new data is said to generalize well.

A machine learning practitioner will want to have confidence in the model's ability to generalize before deploying the model

to make predictions in real-time, and will not necessarily have a new dataset of previously unseen observations to run through the model. To get around this, the original dataset is often split into two parts. The first part of the dataset is referred to as the training set and is used to determine the coefficient weights. The second part of the dataset is referred to as the test set, and this set is run through the model derived from the training set. The accuracy of the predictions on the test set is compared to the accuracy of the predictions on the training set to determine the extent to which the model generalizes [4].

A model that has high training accuracy, but low test accuracy, is said to be overfitting the data. This means that the model, in its efforts to minimize  $\epsilon$ , has become too complex and focuses too closely on the samples in the training dataset. By chasing patterns in the training data caused more so by random chance than by the true characteristics of  $x$ , the model no longer generalizes to the unseen samples in the test set [6][4]. An overfit model describes characteristics in the training data that are not in the test data, leading to poor predictions on the test set.

A model can also underfit the data, which means the model is failing to capture the relationship between  $Y$  and  $x$  and will likely perform poorly on both the training and test datasets.

**2.1.3 The Bias/Variance Trade-off.** Bias and Variance are two important components related to training models using machine learning. Variance describes the extent to which a model changes due to small adjustments in the training data. Since the training data used to fit a model can vary, it is reasonable to expect that a model will change when different samples are selected into the training dataset, but ideally the model changes only slightly [6]. If a model is quite complex and is overfitting the training data, then slight changes in the training samples can have a large effect on the coefficient weights. Low variance is preferable [6].

Bias refers to the error that occurs when trying to describe a phenomenon using a model. For example, if a machine learning technique assumes a linear relationship between the independent and dependent variables, but the relationship is highly non-linear, then the model has high bias [6]. A model with high bias will make many erroneous predictions because the estimated relationship between  $x$  and  $Y$  is not closely aligned with the actual relationship between  $x$  and  $Y$ .

As a model becomes more complex and able to fit to the perceived important information in the training data, variance will increase and bias will decrease. The model will become more flexible and therefore more sensitive to variations in the training data, but will reduce bias by better estimation of the relationship between  $x$  and  $Y$ , resulting in a reduction in the prediction error. The important part of the relationship between these two components is that as a model becomes more complex, the bias decreases more rapidly than the variance increases, so the trade-off of increasing variance while decreasing bias leads to a net gain in improvement of the model [6]. However, there is a point at which the model becomes too complex and the net gain begins to disappear. Increased model complexity leads to significantly higher variance without appreciable improvement in bias [6].

**2.1.4 Model Evaluation.** Several statistics can be used for evaluating model accuracy. For classification problems, a basic technique for evaluation is the confusion matrix.

$$\begin{Bmatrix} TN & FP \\ FN & TP \end{Bmatrix}$$

This is the general framework of a confusion matrix which shows the counts of each type of prediction and the accuracy of that prediction. A true positive (TP) is an outcome that is predicted to be positive and is positive in reality [2]. A true negative (TN) is an outcome that is predicted to be negative and is negative in reality [2]. These are the preferred responses. In the context of hospital readmissions, a true positive is a prediction that a patient in the test dataset, according to the trained model, will be readmitted to the hospital within 30 days, and this occurs in reality. A true negative is a prediction that a patient in the test dataset will not be readmitted, and this occurs in reality.

On the other hand, a false positive (FP) is an outcome that is predicted to be positive but is negative in reality [2]. A false negative (FN) is an outcome that is predicted to be negative but is positive in reality. These are errors in prediction [2]. If a healthcare provider acts on a false positive, that could mean that a patient, who without intervention would not have been readmitted within 30 days, received resources and attention that were not necessary. In the case of a false negative, this means a patient who eventually did get readmitted within 30 days, but was said to be of low-risk of readmission, could have benefited from additional attention and resources from a healthcare team.

These four components - true positives, true negative, false positives, and false negatives - can be combined to create more nuanced metrics. Two of those metrics are sensitivity and specificity. Sensitivity refers to the true positive detection rate. This is the percentage of positive occurrences that are successfully identified [2]. Specificity is the true negative detection rate. This is the percentage of negative occurrences that are successfully identified [2].

In the context of readmissions, low sensitivity means many patients who eventually get readmitted are not predicted to be high-risk before the readmission occurs. Low specificity means that many patients who would not otherwise be readmitted are predicted to be readmitted. There is a trade-off between sensitivity and specificity, and an improvement in one often causes the other to worsen. Preference toward sensitivity or specificity often depends on the cost of incorrect predictions.

A patient who otherwise would not be readmitted who is predicted to be high-risk is the type of case that will incur unnecessary resources. While this requires healthcare providers to invest resources that are not needed, the readmission is nevertheless avoided and there are potentially other benefits achieved by the hospital, such as increased satisfaction of the patient and their family. On the other hand, a patient who eventually gets readmitted but was not identified beforehand will likely be costly to a hospital in a couple ways. The provider must dedicate resources to stabilizing and healing the patient, while also incurring penalties if this type of readmission occurs frequently. If the expense of an unexpected readmission is higher than the expense of deploying unnecessary resources to low-risk patients, then a model that favors higher sensitivity at the expense of lower specificity is preferable.

Sensitivity and specificity can be assessed in tandem by the receiver operating characteristic (ROC) curve, which is quite useful for evaluating supervised classification models. The ROC curve plots the true positive rate against the false positive rate (100 minus the true negative rate) for varying decision thresholds. This illustrates the trade-off between sensitivity and specificity and can provide guidance on which decision threshold is appropriate for the task [2]. ROC curves are often leveraged to evaluate the performance of models by calculating the area under the ROC curve, also known as the AUC. The goal is to maximize the AUC value, and that value points to the optimal balance between sensitivity and specificity [2].

## 2.2 Logistic Regression

**2.2.1 Logistic Regression - Intuition.** Logistic regression models the probability that a sample belongs to a certain class given the feature values of the sample [4]. This probability can be represented as:

$$p(x) = Pr(Y = 1|X)$$

In the context of predicting hospital readmissions, this translates to the likelihood that a patient will be readmitted within 30 days of discharge given the patient's characteristics. To determine the probability, logistic regression utilizes the logistic function, which takes in the coefficient weights and feature responses for each sample and returns a the probability - a number between 0 and 1 [4]. In the case of logistic regression involving multiple features, the model takes the form:

$$f(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

The model is fit to the data by adjusting the coefficient weights using a method called maximum likelihood. The intuition of this process is that the estimates for the coefficients are set such that the predicted probability of a certain outcome corresponds as closely as possible to the actual label of that sample. This means that the ideal coefficient weights, when plugged into the logistic function, return a number close to one for the readmitted patients and a number close to zero for the patients not readmitted [4].

**2.2.2 Logistic Regression - Data Pre-processing.** Data often need to be processed prior to using logistic regression because this machine learning technique requires numerical data. The diabetes dataset contains a combination of continuous and categorical features. For example, 'num\_procedures' and 'num\_lab\_procedures' are continuous features that describe the number of procedures and the number of lab procedures, respectively. Since these columns already contain numerical data, these features are ready to use as-is. Other columns such as 'A1Cresult' includes values such as 'A1Cresult\_>7', 'A1Cresult\_>8', 'A1Cresult\_None' and 'A1Cresult\_Norm'. The first issue is that this features is represented by text values, which will not work with logistic regression. These values must be encoded to work properly. If a categorical feature with four unique values, or levels, has an ordinal scale, the text values can be encoded as sequential numbers, such as 1, 2, 3 and 4. If a categorical features with four levels has a nominal scale, as is the case with the feature

'A1Cresult', an effective encoding strategy is to create one dummy column for each level in the original categorical column.

The Python library pandas has a function called 'get\_dummies' that will create one column for each level in a categorical column, and each of those dummy columns will only contain 0's and 1's. In the case of the column 'A1Cresult', this process will yield 4 columns. For each observation, a 1 will appear in the column corresponding to the value of the original feature. For example, if an observation had a value of 'A1Cresult\_>7', the observation will have a 1 in the 'A1Cresult\_>7' dummy column and 0's in the other three A1C dummy columns. This process is repeated for all nominal categorical variables.

Three categorical columns in the dataset have several hundred unique values, which can be problematic. The columns 'diag\_1', 'diag\_2' and 'diag\_3' have 695, 724 and 757 unique values describing ICD9 diagnosis codes, respectively. The first diagnosis column is considered to be the primary diagnosis of the stay, and 'diag\_2' and 'diag\_3' contain any additional diagnoses documented during the stay. Running these columns through the 'get\_dummies' procedure would yield a total of 2,176 dummy columns, which would greatly increase the dimensionality of the dataset. Further, many ICD9 codes are used only a few times in the dataset, which means it is quite likely that, depending on how the training and test data is split, all observations of a particular code only fall in either the training set or the test set.

One solution to this problem is to 'bin' the information into categories. Each ICD9 code belongs to a category. For example, ICD9 code '250.62 - Diabetes with neurological manifestations, Type II, uncontrolled' is in the ICD9 category 'Endocrine, Nutritional, Metabolic, Immunity'. Each ICD9 code can be binned into one of 19 categories. Further, instead of having three columns for each ICD9 category (because each unique ICD9 code can appear in any of the three diagnosis columns), the data can be processed such that there is one column for each ICD9 category, and each observation can have up to three 1's in these 19 dummy columns. This loses the distinction between primary and secondary diagnoses, but reduces computation time and reduces the likelihood of the rare diagnosis codes only appearing in the test set or training set.

Another column in the dataset called 'medical\_specialty' has a high number of unique values with 71 different responses, and also is null in nearly half of the observations. Rather than turning this feature into 71 different dummy variable columns, it is noted that there is redundancy between the 'medical\_specialty' and the diagnosis code columns. For example, if a patient has a diagnosis code in the 'Pregnancy, Childbirth, and the Puerperium' category, they are often in the obstetrics medical specialty. Given this redundancy, the high percentage of null values and in the interest of reducing the complexity of the dataset, the 'medical\_specialty' column is not included in the final dataset.

Several patients have multiple observations captured in the dataset. Logistic regression requires that the observations be independent, so including multiple inpatient encounter for individual patients violates this requirements. To solve this problem, the initial count of 101,766 observations is reduced down to 69,988 observations by keeping only the first encounter for each 'patient\_nbr'. The first encounter per patient is considered to be the observation with the lowest 'encounter\_id', which operates on the assumption that

IDs are incremented by 1 and allocated sequentially as inpatient admissions occur.

Lastly, the response label in the original dataset is represented with three levels and is described in text. The column ‘readmitted’ contain the values ‘NO’, ‘>30’ and ‘<30’. Since observations with the label ‘>30’ days were not readmitted within thirty days, these labels were converted to ‘NO’. The remaining responses of ‘NO’ and ‘<30’ were encoded as 0 and 1, respectively.

**2.2.3 Logistic Regression - Data Quality Evaluation.** When creating dummy columns, whether through simple methods, such as the ‘A1Cresult’ transformation, or more complex methods, such as the ICD9 diagnosis binning transformation, special consideration must be given to collinearity and multicollinearity between features. For example, if a feature called ‘gender’ contains two values, male and female, and this feature is converted into two dummy features, these two features will be collinear. Where one feature column has a value of one, the other will have a zero, and visa versa. This means that one feature column can perfectly predict the value of the other feature column. We only need the female column to know if the observation pertains to a male or female, so the inclusion of the male column would be redundant. This is problematic for the model because the two feature columns provide an identical explanation of the variance in the dependent variable, and neither adds additional value while in the presence of the other. When this issue manifests between two columns, this means the columns are collinear. Multicollinearity refers to a situation where this redundancy occurs between three or more columns. If the combination of three columns explains most of the variation explained by another single column, then there is multicollinearity in the data.

Collinearity and multicollinearity increase the variance of the coefficient weights, which would make the model very sensitive to changes in the training data. This instability of the weights means that it can be difficult to decide which predictors have a high influence on the outcome, and can even cause the sign of the coefficient to change [3]. Under stable conditions, a positive coefficient can be interpreted to mean that the associated feature contributes to a higher probability of readmission, and a negative coefficient can be interpreted to mean that the associated feature contributes to a low probability of readmission [6]. The instability that multicollinearity creates in the coefficient weights make it dubious to make inferences from the signs of the weights.

Datasets with collinearity and multicollinearity issues are considered to be ill-conditioned, which will reduce the ability to create a meaningful model with the data. Problematic features need to be strategically identified and removed. A dataset can be evaluated for problems using several linear algebra methods. The matrix rank is a single value that can give an overall assessment of the relationship between features. In a dataset, which can be represented as a matrix, that has more rows than columns, the ideal matrix rank value is equal to the number of columns. When the matrix rank value is equal to the number of columns this means the matrix is considered to be full rank [12]. A full rank matrix contains only linearly independent features. On the other hand, if a feature in a dataset is linearly dependent, then the rank of the matrix is reduced. For example, if we were to keep both the male and female gender

dummy columns, these features would be considered linearly dependent, and would therefore reduce the rank of the matrix. Each linearly dependent feature in a dataset reduces the matrix rank.

A correlation matrix provides a correlation statistic for each pair of variables. The values fall between -1 and 1, and the closer the value is to -1 or 1, the stronger the relationship between the two variables. Features with high correlation are considered to be collinear. This technique is effective at finding collinearity, but is not well-suited to finding multicollinearity because the correlation matrix only shows the relationship between pairs of variables.

The correlation matrix can also be used to find the determinant of the dataset. The determinant is a single value and will reveal if there are any highly or perfectly correlated columns, which suggests there is collinearity among features. The determinant value ranges between 0 and 1. A value of zero means the correlation matrix is singular. In other words, the correlation matrix contains at least one pair of perfectly correlated features. A near-zero determinant value means there is one or more pair of features that is nearly correlated. A higher determinant value is preferable.

These methods are effective at describing the overall health of the dataset and simple relationships between pairs of features. To find multicollinearity, more nuanced techniques need to be deployed. One approach is to determine the variance inflation factor (VIF) for each independent variable. The VIF measures the increase of variance in the coefficient estimates that is caused by the inclusion of a particular variable [8]. This technique fits each independent variable, one at a time, against all of the other independent variables. This can be represented by the following sequence of equations:

$$\begin{aligned} X_1 &= \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \dots + \beta_kX_k \\ X_2 &= \beta_1X_1 + \beta_3X_3 + \beta_4X_4 + \dots + \beta_kX_k \\ X_3 &= \beta_1X_1 + \beta_2X_2 + \beta_4X_4 + \dots + \beta_kX_k \\ &\dots \\ X_k &= \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \dots + \beta_{k-1}X_{k-1} \end{aligned}$$

For a dataset with k-features, the dataset is fit k-times, once for each independent feature. The VIF for each feature is calculated by the equation:

$$VIF_k = \frac{1}{1 - R_k^2}$$

Each fitted model has an  $R^2$ , which is the coefficient of determination, or R-squared, and it describes the proportion of variation in the ‘dependent’ variable that is described by the independent variables. A high R-squared means that the independent variables explain a significant amount of variation in the dependent variable. In the context of VIF, if one independent variable is thoroughly explained by the other independent variables, the R-squared will be high which will lead to a high VIF. While the threshold for acceptable VIF values differs, the documentation for the Python library statsmodels recommends using a threshold of 5 [8]. To achieve a value of 5 or less, the  $R^2$  for an independent variable must 0.80 or less. In other words, the independent variable being considered by the VIF method must be less than 80% explained by the other independent variables.

The elimination of problematic variables in this dataset is handled by a custom Python function that identifies the features with

the highest VIF and selectively removes those features from the dataset. The Python functions works by calculating the VIF for each independent variable. It then iterates through each group of dummy columns that stemmed from a single categorical column, and, for each group, deletes the column with the highest VIF value if that value is above the threshold. The function also removes ratio-scaled features, such as ‘num\_procedures’, that have a high VIF value. This whole process is looped until zero features have a VIF above the threshold.

Before trimming features based on VIF, the dataset included 175 features, with a matrix rank of 150 and a correlation matrix determinant of zero, meaning the coefficient matrix was singular. This means there were several linearly dependent features and at least one pair of perfectly correlated features. After trimming 52 features based on VIF, the dataset includes 123 features with a full rank of 123 and a correlation matrix determinant of 0.00697. While this determinant value is still relatively low, the determinant increased over each iteration of the Python function from 0.0 to 2.85e-26 to 0.0056 to 0.00697, representing several orders of magnitude in improvement from the originally singular matrix. Most importantly, the matrix now has full rank with 0 linearly dependent features.

**2.2.4 Logistic Regression - Feature Selection.** With the issue related to multicollinearity among the independent variables largely resolved, the coefficient weights of the model will be more stable, allowing for inferences to be made based on the sign and magnitude of the weights. The next step is to strategically choose which features to use when training the model. Recursive feature selection (RFE) is one strategy for choosing which features have the highest significance in predicting the likelihood of readmission within 30 days. The intuition behind RFE is that it repeatedly fits the model on the training data. The first iteration includes all features, and each subsequent fitting of the data drops the least significant feature or features from the previous iteration. Python has a library called scikit-learn which includes tools to execute RFE on a dataset. The user may choose how many features to trim after each iteration, as well as choose how many features the final model should have. The process will repeatedly fit the narrowing set of features to the data until the preferred number of the most important features is reached.

There is an extension of RFE called RFECV, which helps to determine the ideal number of features. When using RFE on its own, the user must arbitrarily choose the preferred number of procedures. RFECV functions by calculating the accuracy of the model after each iteration of trimming features and re-fitting the model. The number of features used at the step in which the model performance is best is determined to be the ideal number of features to use. The ‘transform’ method of RFECV will then trim down the original dataset to the selected features. After running RFECV on the remaining 123 features, the process selected 57 features that led to the highest accuracy rate.

**2.2.5 Logistic Regression - Execute Analysis.** The set of independent variables is trimmed down further to the 57 features selected by RFECV as being the most important for predicting likelihood of readmission within 30 days. The next step is to train the logistic regression model, and then test the accuracy of the model. Scikit-learn has a function that randomly splits the dataset into training

and test sets, and also allows the user to decide the size of the test dataset in terms of proportion of overall data. After splitting the features and labels into training and test sets, the data is ready for fitting.

Scikit-learn also has a process for executing logistic regression, and there is a parameter that controls the way the algorithm minimizes coefficients. The default setting is L2 regularization, which determines coefficients that can approach zero (meaning the associated feature does not have a large effect on the outcome) but never fully reach zero. This regularization of coefficients effectively determines how much effect each feature has on the prediction. The less significant features will have a coefficient close to zero. L1 regularization is another option, which sets the less significant features to exactly zero, which can be viewed as another form of feature selection [4].

There is another parameter called C, which dictates the strength of the regularization. Higher values of C lead to less regularization. This means that a model trained with a high value of C will value fitting each observation as closely as possible, whereas a lower value of C will train the model in a way that tries to fit the data more generally [4]. A high value of C will lead to higher weight values, and a low value of C will lead to weights that are much closer to zero.

**2.2.6 Logistic Regression - Evaluate Analysis.** The model is trained using both L1 and L2 regularization, and each regularization type is fit using three different values for C: 0.01, 1.0 and 100.0. Figure 1 shows the coefficient weights using L2 regularization and the three different values of C. It is evident that higher values of C lead to larger weights. Figure 2 show the coefficient weights using L1 regularization, again with the different values of C. In addition to the observation that higher value of C lead to larger weights, it is also interesting to note that using 0.01 for the value of C sets all but four weights equal to zero. The four features chosen by this model are the numbers of inpatient encounters, age 50-60, transferred to a skilled nursing facility and discharged to another rehabilitation facility. This pair of L1 regularization and 0.01 for C has the highest training and test set accuracy. The training accuracy is 91.063% and the test accuracy rate is 90.0827%. In the original dataset of the 69,998 observations, 63,704 were not readmitted. This is a rate of 91.02%. This is only slightly smaller than the training accuracy and larger than the test accuracy, which means the model performs closely to the rate that would be achieved if a person guessed that every case would not be readmitted. The confusion matrix for L1, C = 0.01 model is:

$$\begin{Bmatrix} 12713 & 2 \\ 1282 & 1 \end{Bmatrix}$$

12,713 true negatives were identified and 1 true positive, for a total of 12,714 accurate predictions. There were 2 false positives and 1,282 false negatives. The model is effective at predicting patients who will not be readmitted, but the high number of false negatives, compared to the extremely low count of true negatives, demonstrates that the model is not performing well at identifying patients who eventually get readmitted. The models with C values of 1.0 and 100.0 have a true negative detection count of 4, slightly

higher than the 1 observation classified correctly by the L1, C = 0.01 model.

The relationship between the true positive and false positive rates can be visualized with an ROC curve. Figure 3 show the ROC curve for the L1, C = 0.01 model. The black dotted line represents the 50/50 chance curve, which is equivalent with guessing. The ROC curve extends slightly above the 50/50 chance curve, which means the predictive power is slightly higher than random guessing. This is described by the AUC, which has a value of 0.50013. This is consistent with the conclusion that model is only slightly better than chance. Figure 4 shows the ROC curve for the L1, 100 model, and the ROC curve bends further away from the 50/50 chance curve, and the AUC is slightly higher at 0.5013. This is consistent with the observation that the model with the higher value of C has a higher true negative detection rate. Ideally, the ROC curve is as close to the upper left hand corner as possible, which would represent a high true positive rate with a low false positive rate.

### 3 CONCLUSION

The predictive power of the logistic regression model chosen for this analysis appears to be slightly better than random guessing, but not significantly better. The high proportion of false negatives means many patients who are at high risk of readmission within 30 days, and later get readmitted, are not being identified by the model. This is a domain where high sensitivity is favored over high specificity, but the model conversely has low sensitivity and high specificity. To improve the predictive power of the model, it might be helpful to include features that have more to do with behavioral and social characteristics, as well as socioeconomic indicators. Attributes such as literacy, obesity, annual income, smoking status, medication regimen adherence, utilization of family and community support and employment status are a few features that come to mind that may lend to better explaining the likelihood of readmission within thirty days. Features of this type may help describe the extent to which a patient is able to manage his or her own care outside of the hospital. Patients who cannot read or who do not adhere to the recommended medication regimen, for example, are patients who can reasonably be said to be less capable of providing consistent and effective care to themselves in the home setting. Attributes such as this are not available in the dataset, but common sense suggests this information would be helpful.

Further, logistic regression is just one type of machine learning technique capable of performing classification. Support vector machines and decision trees are two other techniques that would be worth exploring to see if modeling the data using different machine learning algorithms improves the sensitivity of the model.

### A ACCOMPANYING JUPYTER NOTEBOOK AND REQUIREMENTS

The accompanying Jupyter Notebook is available at: <https://github.com/bigdata-i523/hid331/blob/master/project/project.ipynb>

The requirement file is available at: <https://github.com/bigdata-i523/hid331/blob/master/project/requirements.txt>

### ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and his teaching assistants for their support throughout the semester.

### REFERENCES

- [1] Cristina Boccuti and Gisella Casillas. 2017. Aiming for Fewer Hospital U-turns: The Medicare Hospital Readmissions Reduction Program. Online. (March 2017). <http://files.kff.org/attachment/Issue-Brief-Fewer-Hospital-U-turns-The-Medicare-Hospital-Readmission-Reduction-Program>
- [2] Christopher M Florkowski. 2008. Sensitivity, Specificity, Receiver Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. *Clinical Biochemistry Review* 29 (August 2008), S83–S87. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556590/>
- [3] Jim Frost. 2013. What Are the Effect of Multicollinearity and When Can I Ignore Them? Online. (May 2013). <http://blog.minitab.com/blog/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them>
- [4] Sarah Guido and Andreas Müller. 2017. *Introduction to Machine Learning with Python* (1st edition ed.). O'Reilly Media, 1005 Gravenstein Highway North, Sebastopol, CA, 95472.
- [5] Danning He, Simon C Mathews, Anthony N Kalloo, and Susan Hufless. 2013. Mining High-dimensional Administrative Claims Data to Predict Early Hospital Readmissions. *Journal of Informatics in Health and Biomedicine* 21, 2 (March 2013), 272–279. <https://doi.org/10.1136/amiajnl-2013-002151>
- [6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning*. Springer Science and Business Media, 11 W 42nd St, New York, NY, 10036. <https://doi.org/10.1007/978-1-4614-7138-7>
- [7] Paul LaBrec. 2016. Analyze this! Administrative claims data or EHR data in health services research? Online. (January 2016). <https://www.3mhisinsideangle.com/blog-post/analyze-this-administrative-claims-data-or-ehr-data-in-health-services-research/>
- [8] Josef Perktold, Skipper Seabold, and Jonathan Taylor. 2012. Source code for statsmodels.stats.outliers\_influence. Online. (January 2012). [http://www.statsmodels.org/dev/\\_modules/statsmodels/stats/outliers\\_influence.html#variance\\_inflation\\_factor](http://www.statsmodels.org/dev/_modules/statsmodels/stats/outliers_influence.html#variance_inflation_factor)
- [9] Jordan Rau. 2016. Medicare's Readmission Penalties Hit New High. Online. (August 2016). <https://khn.org/news/more-than-half-of-hospitals-to-be-penalized-for-excess-readmissions/amp/>
- [10] Khader Shameer, Kipp W Johnson, Alexandre Yahia, Riccardo Miotto, Li Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P Sengupta, Annette Gelijns, Alan Moskowitz, Bruce Darro, David Reich, Andrew Kasarskis, Nicholas P Tatone, Sean Pinney, and Joel T Dudley. 2016. Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort. In *PSB, Pacific Symposium on Biocomputing* (Ed.), Vol. 22. Pacific Symposium on Biocomputing, Pacific Symposium on Biocomputing, 1 N Kaniku Dr, Waimea, HI, 96743, 276–287. <https://www.ncbi.nlm.nih.gov/pubmed/27896982>
- [11] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International* 2014 (April 2014), 1–11. <https://doi.org/10.1155/2014/781670>
- [12] Stat Trek. 2017. Matrix Rank. Online. (2017). <http://stattrek.com/matrix-algebra/matrix-rank.aspx>

### B FIGURES

[Figure 1 about here.]

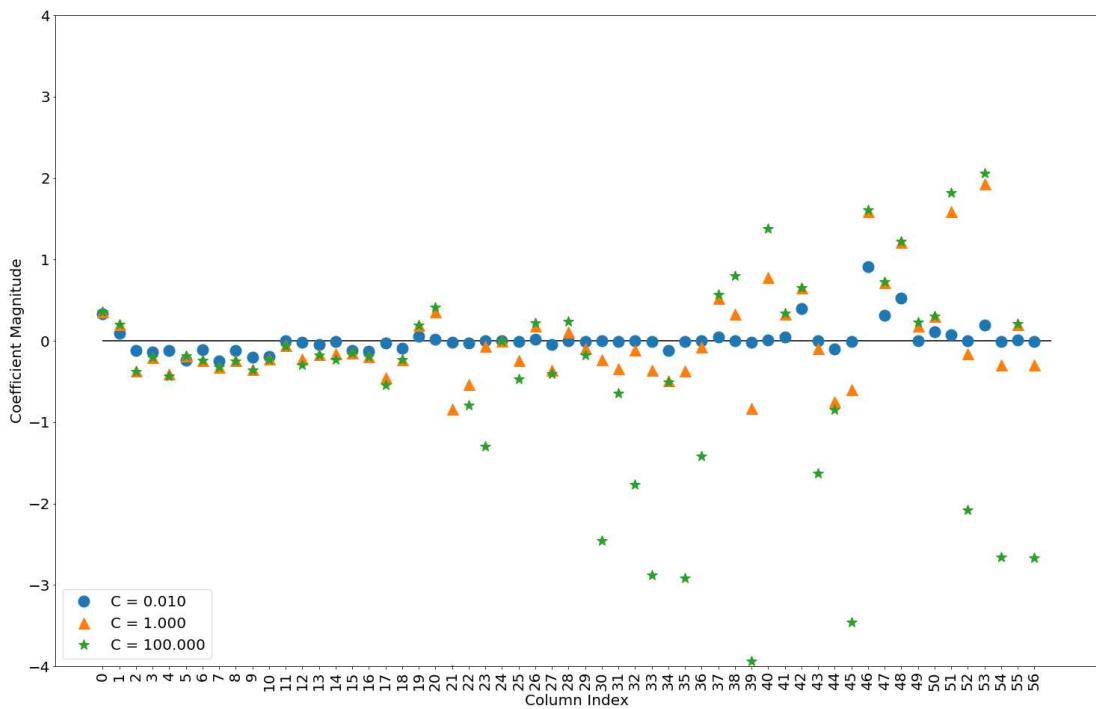
[Figure 2 about here.]

[Figure 3 about here.]

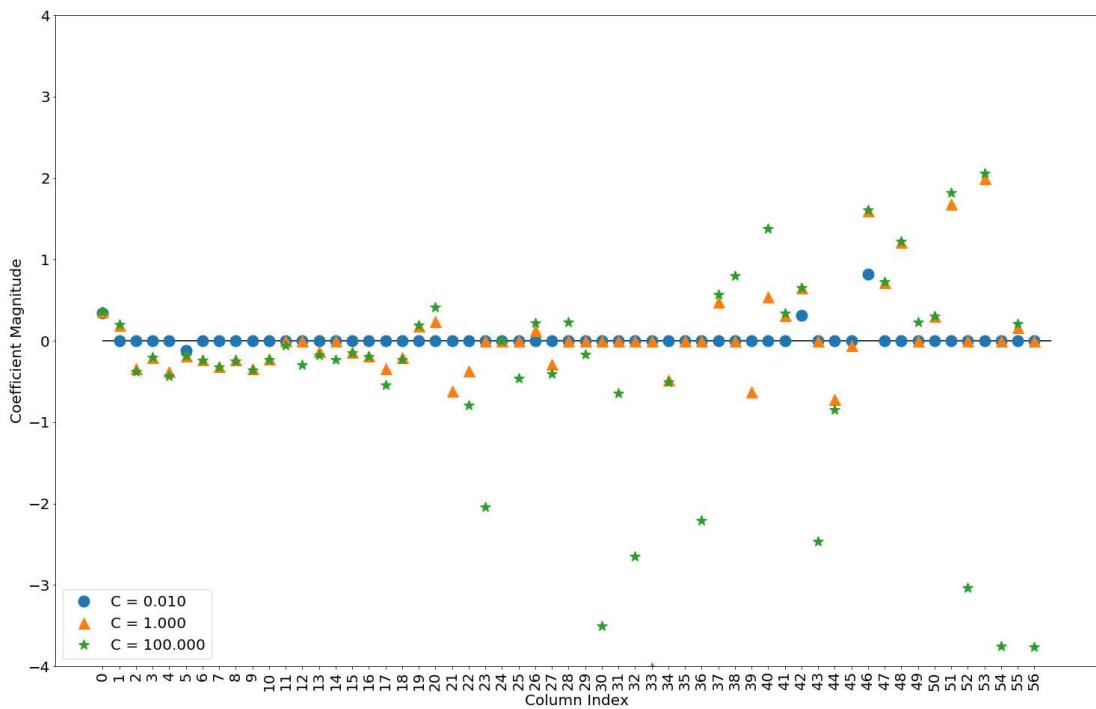
[Figure 4 about here.]

LIST OF FIGURES

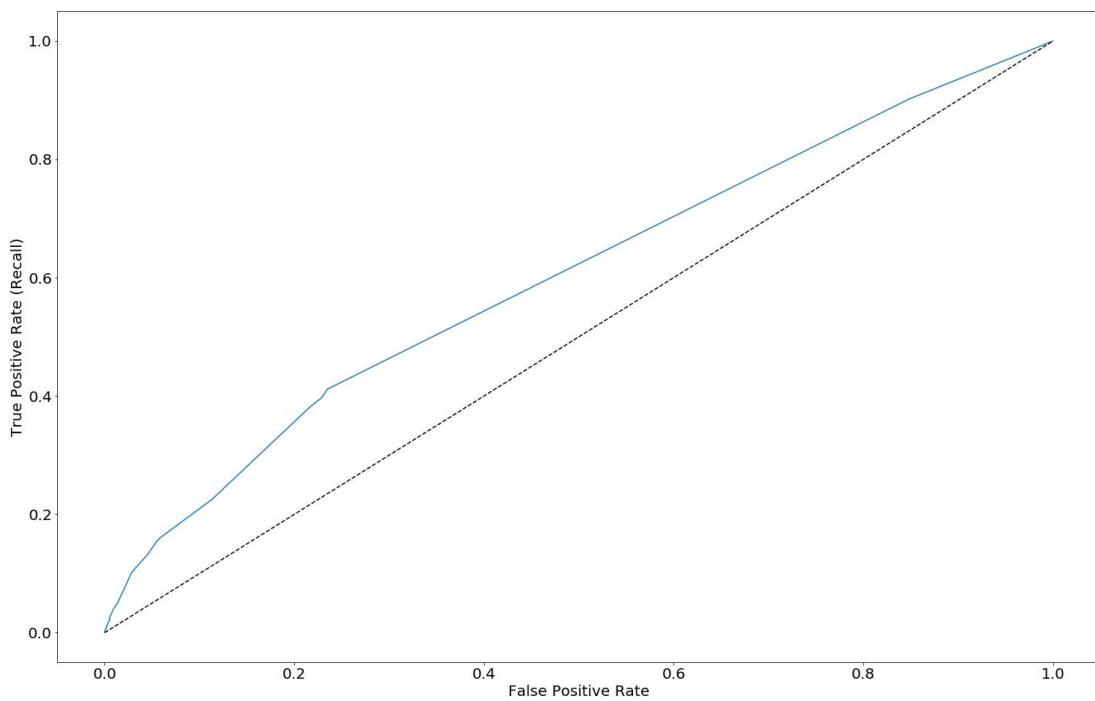
1	Logistic Regression Weights By C-Value, L2 Regularization	9
2	Logistic Regression Weights By C-Value, L1 Regularization	10
3	ROC Curve, L1 Regularization, C = 0.01	11
4	ROC Curve, L1 Regularization, C = 100.0	12



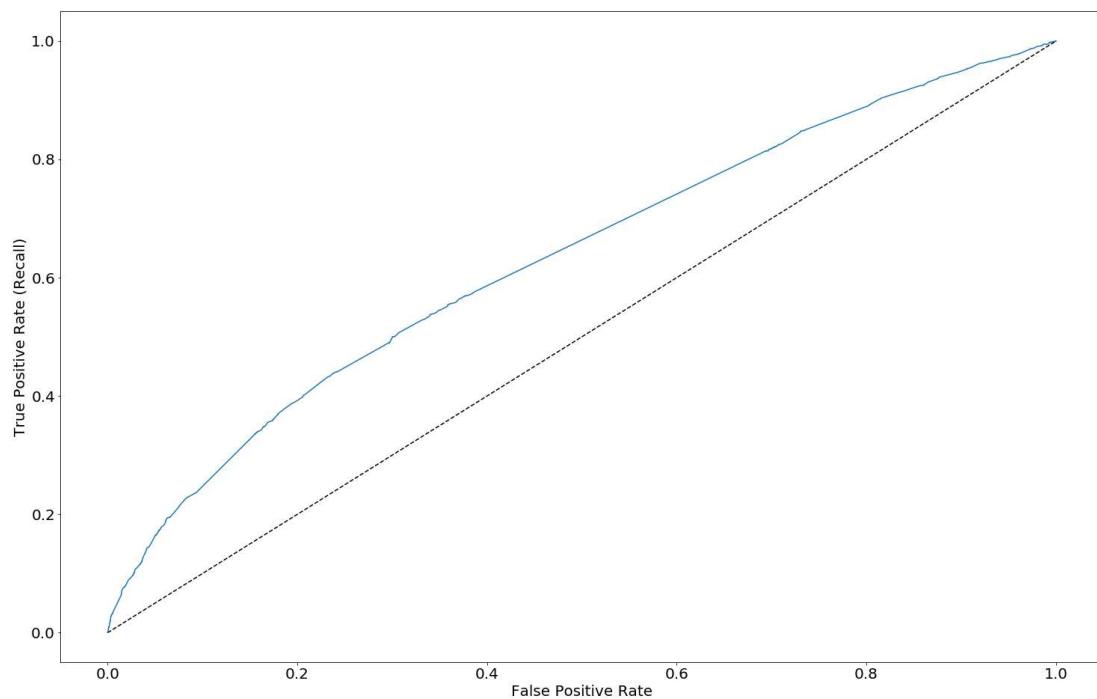
**Figure 1: Logistic Regression Weights By C-Value, L2 Regularization**



**Figure 2: Logistic Regression Weights By C-Value, L1 Regularization**



**Figure 3: ROC Curve, L1 Regularization, C = 0.01**



**Figure 4: ROC Curve, L1 Regularization, C = 100.0**

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-12-16 09.38.18] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.8s.
```

---

```
Compliance Report
```

---

```
name: Tyler Peterson
hid: 331
paper1: Oct 22 17 100%
paper2: Nov 6 17 100%
project: Dec 4 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

```
-----  
12  
wc 331 project 12 6935 report.tex  
wc 331 project 12 7039 report.pdf  
wc 331 project 12 906 report.bib  
  
find "  
-----  
  
passed: True  
  
find footnote  
-----  
  
passed: True  
  
find input{format/i523}  
-----  
  
6: \input{format/i523}  
  
passed: True  
  
find input{format/final}  
-----  
  
passed: False  
  
floats  
-----  
  
198: The model is trained using both L1 and L2 regularization, and  
each regularization type is fit using three different values for  
C: 0.01, 1.0 and 100.0. Figure \ref{f:weightsl2} shows the  
coefficient weights using L2 regularization and the three  
different values of C. It is evident that higher values of C lead  
to larger weights. Figure \ref{f:weightsl1} show the coefficient  
weights using L1 regularization, again with the different values  
of C. In addition to the observation that higher value of C lead  
to larger weights, it is also interesting to note that using 0.01  
for the value of C sets all but four weights equal to zero. The  
four features chosen by this model are the numbers of inpatient  
encounters, age 50-60, transferred to a skilled nursing facility  
and discharged to another rehabilitation facility. This pair of  
L1 regularization and 0.01 for C has the highest training and
```

test set accuracy. The training accuracy is 91.063\% and the test accuracy rate is 90.0827\%. In the original dataset of the 69,998 observations, 63,704 were not readmitted. This is a rate of 91.02\%. This is only slightly smaller than the training accuracy and larger than the test accuracy, which means the model performs closely to the rate that would be achieved if a person guessed that every case would not be readmitted.

- 212: The relationship between the true positive and false positive rates can be visualized with an ROC curve. Figure \ref{f:roccurve001} show the ROC curve for the L1, C = 0.01 model. The black dotted line represents the 50/50 chance curve, which is equivalent with guessing. The ROC curve extends slightly above the 50/50 chance curve, which means the predictive power is slightly higher than random guessing. This is described by the AUC, which has a value of 0.50013. This is consistent with the conclusion that model is only slightly better than chance. Figure \ref{f:roccurve100} shows the ROC curve for the L1, 100 model, and the ROC curve bends further away from the 50/50 chance curve, and the AUC is slightly higher at 0.5013. This is consistent with the observation that the model with the higher value of C has a higher true negative detection rate. Ideally, the ROC curve is as close to the upper left hand corner as possible, which would represent a high true positive rate with a low false positive rate.
- 238: \begin{figure}[!ht]
- 239: \centering\includegraphics[width=\columnwidth]{images/weightsl2.png}
- 240: \caption{Logistic Regression Weights By C-Value, L2 Regularization}\label{f:weightsl2}
- 243: \begin{figure}[!ht]
- 244: \centering\includegraphics[width=\columnwidth]{images/weightsl1.png}
- 245: \caption{Logistic Regression Weights By C-Value, L1 Regularization}\label{f:weightsl1}
- 248: \begin{figure}[!ht]
- 249: \centering\includegraphics[width=\columnwidth]{images/roccurve001.png}
- 250: \caption{ROC Curve, L1 Regularization, C = 0.01}\label{f:roccurve001}
- 253: \begin{figure}[!ht]
- 254: \centering\includegraphics[width=\columnwidth]{images/roccurve100.png}
- 255: \caption{ROC Curve, L1 Regularization, C = 100.0}\label{f:roccurve100}

figures 4

```
tables 0
includegraphics 4
labels 4
refs 2
floats 4
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====  
The following tests are optional  
=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# **Big Data Analytics Role in Reducing Healthcare Costs in the United States**

Judy Phillips

Indiana University

PO BOX 4822

Bloomington, Indiana 47408

judkphil@iu.edu

## **ABSTRACT**

In the United States more money is spent on health care than in any other industrialized country in the world. Yet, health care access is often problematic and health care quality indicators are lower or mediocre as compared to other countries with similar economic status. Insights offered by Big Data Analytics can find solutions that will significantly lower costs and improve delivery of health care in the United States. These solutions have the potential to save billions of dollars in health care costs and to improve the quality of care for millions of Americans.

## **KEYWORDS**

I523, HID332, health care costs, predictive analytics, electronic health records, big data

## **1 INTRODUCTION**

Health care spending in the United States greatly exceeds the spending of other industrialized countries. Americans spend 3 trillion dollars annually on health care. Health expenditures currently account for 17.6 percent of the Gross National Product (GDP) and are expected to increase at an average rate of 5.8 percent through 2025. Health care spending has exceeded growth of the Gross National Product (GDP) in 42 of the previous 50 years [2]. Health spending threatens the nation's fiscal health [29]. Despite the excessive spending, the United States ranks among the worst on measures of health care quality, health access equity, and quality of life [22]. Policy makers do not know how to respond.

Big data analytics has the potential to help manage and address some of the cost issues while simultaneously improving patient health outcomes. Big Data ability gives us the ability to combine and analyze data from a wide variety of sources in ways that have never before been possible. This new information is providing new and valuable insights into ways to provide more effective and efficient patient care. The associations, patterns, and trends in big data may hold the key to reducing expenditures, improving care, and saving lives [29]. The information is being used to achieve more accurate and timely diagnoses, better match treatment plans to patient needs, and predict and identify at-risk patients and populations [22]. Mobile applications are being used to monitor patient care in real time. Big data can reduce health care waste, improve coordination of care, expose fraud and abuse, and to speed up the research and development pipeline.

The cost savings estimates are substantial. McKinsey and Company estimates that Big data analytics has the potential to reduce health care costs in the United States by 12 to 17 percent. This

equals to a savings of between 348 to 493 billion dollars annually [6].

Some of the tools and methodologies that big data uses to introduce efficiencies into the American health care system include: Outcome based reimbursement methodologies, electronic health records, medical device monitoring, predictive analytics, evidence based medicine, genomic analysis, and claim prepayment fraud analysis. Big data technologies are adding value and improving efficiency in almost every area of health care including clinical decision support, administration, pharmaceutical research and development, and population health management.

## **2 COMPARISON TO OTHER COUNTRIES**

According to the Organization for Economic Cooperation and Development (OECD), the United States spends 2.5 times per person than the average of OECD related industrialized nations. In 2016, the United States spent 9822 dollars per person annually on health care. In comparison, the average amount spent per person among all OECD nations was 4033 dollars. The next highest spender was Switzerland at 7919 dollars per person [28]. The average spending as a percentage of Gross National Product (GDP) among OECD nations was 9 percent. Switzerland was again the next highest spender at 12 percent of their Gross National Product (GDP) being spent on health care. According to a McKinsey and Company analysis, the United States spends 600 billion dollars more annually than the estimated benchmark amount as calculated based upon the country's size and wealth as compared to other OECD related nations [18].

The United States lags in many standard indicators of health quality. According to a Commonwealth Fund study of 11 developed countries in 2013, the United States ranked fifth in quality and worst in infant mortality. The United States also ranked last in the prevention of deaths from treatable conditions such as strokes, diabetes, high blood pressure and treatable cancers. The average life expectancy in the United States is 76.3 years. The average life expectancy among all OECD countries is 77.9 years. The incidence of obstetric trauma is 9.6 per 100,000 births in the United States compared to 5.7 incidents per 100,000 in other countries. The statistics for preventable hospital admissions also compare poorly in comparison to other nations. In the United States the hospital admission rate for asthma and COPD was 262 per 100,000 in comparison to the average of 236 per 100,000. Thirty eight percent of the population in the United States is obese. The average obesity rate in other countries is nineteen percent. The United States has fewer physicians and hospitals. In the United States, there are 2.6 practicing physicians and 2.8 hospital beds per 1000 population.

This compares to an average of 3.4 physicians and 4.7 hospital beds on the average in the other countries [28].

The United States has material problems with health care access. Most other OEDC countries have achieved almost universal insurance coverages. On the average, 98 percent of persons in OEDC countries have health insurance. In the United States only 90 percent have health insurance. In addition, cost sharing requirements often make access additionally prohibitive. In 2016, 22.3 percent of the persons in the United States had skipped a medical consultation due to cost concerns. In comparison, the average percentage of individuals who had skipped medical visits due to cost in OEDC nations was 10.5 percent. In the United States 11.6 percent of the population had skipped taking a prescribed medication due to cost in 2016. This compares to an average of 7.1 percent of the population in other OEDC countries who reported foregoing foregone a prescribed medication due to cost [28].

### 3 HEALTH COST DRIVERS

Why is health care so much more expensive in the United States than it is anywhere else in the world? Some of the contributing factors include: the basic health care economic payment structure, inefficient and wasteful use of resources, medical errors, lack of transparency within the system, unnecessary administrative costs, and fraud and abuse.

#### 3.1 Health Care Payment Structure

Many of the cost issues can be contributed to the complex, un-coordinated, multi-payer payment structure. Private insurance companies, Medicaid, and Medicare are the primary payers. An individual's eligibility by payer is dependent upon factors such as employment status, income level, age, and whether or not they are disabled. Most citizens obtain private insurance through their employment. Individuals who are 65 years of age or older or disabled are eligible for Medicare. These individuals may also purchase private Medicare Supplement insurance on their own to pay expenses that Medicare does not cover. Low income individuals may be eligible for Medicaid. If an individual is not eligible for any of these programs, he can purchase individual health insurance from a private insurance company on his own. However, individual health insurance is expensive. According to data from E-care, in 2016, the average monthly premium for an individual was 393 dollars per month. The average cost for family coverage was 1021 dollars per month [39]. In addition, individual insurance policies often include fairly high cost sharing features. Even though subsidies are available through the Affordable Care Act to offset some of these costs, many people choose to forego insurance entirely due to the prohibitive expense.

The system is inefficient and flawed because the basic economic concepts such as supply and demand and competition do not work in this sector. This is because none of the players are incentivized to manage or reduce costs [3]. Consumers do not manage medical utilization because it is being paid for by a third party, the insurance company. Insurance coverage thus insulates patients from the true costs of medical care [3]. Providers are not incentivized to provide efficient, cost effective care. Most providers are paid via a traditional fee for service methodology. That is, providers are

paid for each service that they provide. Traditional fee for service provider payment methodologies that reward health caregivers for quantity instead of quality often result in overutilization of unnecessary tests and treatment procedures. The structure is such that it encourages the production of inefficient and low value services [3]. Insurance companies pass the cost of services on to the consumers in the form of higher premiums year after year. The cost inflation cycle goes on and on.

Administrative waste is another result of the complexity of the United States multi payer payment structure. Each payer has their own rules and standards. Benefit and coverage options can vary dramatically among individuals even within the same insurance company. According to the OEDC 2008 estimates, the United States spends 7.3 percent of health care expenses on administrative activities. This is more than any other country. Comparatively, Germany spends 5.6 percent, Canada spends 2.6 percent and France spends 1.9 percent [28]. Administrative activities include transaction related activities such as billing and claims payment, and regulatory compliance such as those required to comply with government and nongovernment accreditation and regulation including licensing requirements.

#### 3.2 Clinical and Operational Waste

McKinsey and Company estimates that clinical waste amounts to 273 dollars annually [29]. According to the Congressional budget office, 30 percent of United States spending is wasteful or not necessary [8]. There are two types of waste: operational, and clinical [3].

Operational waste results from duplication of services or inefficient production processes. An example would be a duplicate medical service because of lost medical records or the same service already being provided by another caregiver [3].

Clinical waste is created by the creation of low value outputs or care that is not optimally managed. One type of clinical waste is the spending on goods and services that provide marginal or no health benefit over less costly alternatives. Some clinical waste is the result of the uncertainty in the science of medicine. An example would be when a patient is misdiagnosed or when the treatment protocol is uncertain [3]. Other types of clinical waste may be symptoms of a flawed fee for service payment structure. These may include such things as over screening, excessive office visits, or the use of branded instead of generic drugs. Another example is when a newer or more modern treatment is marketed and sold even when it does not provide a better outcome as compared to the traditional treatment. An example was a 2 million dollar prostate cancer machine that was being marketed in 2014. It made the price of the procedure significantly more, but it did nothing to improve the health outcome [8]. Other examples of types of treatment that are the result of clinical waste include avoidable emergency room use, unnecessary hospital admissions, and excessive antibiotic use [3].

#### 3.3 Medical Errors

Medical errors cost the United States system between 17 and 29 billion dollars annually [3]. This amount could be as much as 1 trillion dollars a year if lost productivity is taken into account [27].

This compares to an estimate of 750 million in Canada [3]. The Institute of Medicine estimates that preventable medical errors claim between 44000 and 98000 lives in hospitals each year [3].

### 3.4 Fraud and Abuse

The National Healthcare Anti-Fraud Association estimates losses due to health care fraud at 80 billion dollars annually. Other industry sources estimate fraudulent related losses to be around 200 billion. This accounts for approximately 2 to 3 percent of total health care spending. Research indicates that only 5 percent of these losses are ever recovered [10].

## 4 BIG DATA

Big data refers to electronic data sets that are so large and complex that they cannot be managed with traditional hardware and software. A report delivered to the United States Congress in August 2012 defines big data as large volumes of high velocity, complex, variable data that require progressive techniques and technologies to capture, storage, distribution, manage, and analyze the information. Big data characteristics include variety, velocity, and veracity, and volume [29]. Health care data is big data because it involves the processing of overwhelmingly large complex data sets, from a wide variety of sources and a very rapid speed [29]. In addition, the data is extremely difficult to sort, organize, and decipher [11]. Recent advances in Big Data technology gives us the ability to capture, share and store healthcare data at an unprecedented pace.

### 4.1 Volume

The health care industry has always generated large amounts of data. Data is needed for record keeping, compliance and regulatory reporting and patient care. Historically, this data has been stored in hard copy format. Now, more and more data is being created and stored digitally. In 2011, there were estimated to be 150 Exabytes of health related data. The amount of health related big data is growing rapidly. It is expected to soon reach the zettabyte scale and then soon after that, into the yottabytes [29].

### 4.2 Velocity

Traditionally, health care data has been static: for example, paper files, x-ray films, and prescriptions [29]. Ironically, in many medical situations, the speed of the response can mean the difference between life and death. Increasingly, more and more of the data is being collected in real time and at a rapid pace. For example, medical monitoring devices information collect data continuously, and can support immediate response [29].

### 4.3 Variety

There is an enormous variety of data being collected. The data is in multimedia including images, video, text, numerical, multimedia, paper, and electronic records. Formats include structured, unstructured, and semi-structured. Sources of data include patients, physicians, hospitals, laboratories, research companies, insurance companies, and government agencies. Data comes from web and social media such as Facebook, twitter, health plan websites and smart phone applications. Machine to machine data comes from patient sensors. Biometric data is available such as fingerprints,

genetics, hand writing information and imagining reports [29]. Physicians generate electronic medical records, physician notes, and medical correspondence. Pharmaceutical companies maintain research and development information in medical databases. The United States government houses databases concerning clinical drug trials. Data is collected by the United States Centers Disease Control and Prevention [6].

### 4.4 Veracity

The characteristic Veracity addresses whether the information is credible and error free. Veracity is extremely important in health care because life or death decisions on being based upon the information provided. There is a particular concern because interpretations of unstructured data such as physician notes could be incorrect or imprecise. Big data architecture, platforms, methodologies and tools are designed to take into account the uncertainties of big data analytics [29].

### 4.5 Unstructured Big Data

Unstructured data now makes up about 80 percent of the health care information that is available and is growing exponentially. Sources of unstructured data include: medical devices, physician and nurses notes, and medical correspondence. Being able to access to this information is an invaluable resource for improving patient care and increasing efficiency [22]. Big data technology gives us the ability to capitalize and make use of the valuable clinical information that is unstructured [15].

Traditional databases have well defined structures. The data exists in a table and column format, tables have well defined schemas, and each piece of data is stored within its own well defined space. Big data is not like that at all. Data is extracted from the source systems in its raw format. Massive amounts of this data are stored in a somewhat chaotic fashion in a distributed file system. For example, the Hadoop Distributed File System (HDFS) stores data in directories of files in a hierarchical form. The convention is to store files in 64 Megabyte files in the data nodes using a high degree of compression [15].

Big data is raw data. Big data is not cleansed or transformed in any way. No business rules are applied. The approach is to transform and apply business rules or bind the data semantically as late in the process as possible. In other words, the approach is to bind as close to the application layer as possible [15].

Big unstructured data is less expensive than traditional databases. Most traditional relational databases require propriety software that is associated with expensive licensing and maintenance agreements. Relational databases also need significant specialized resources for design, administration, and maintenance. Because of its unstructured format and open source concept, big unstructured data is much less expensive to own and operate. Big data needs little design work and is easy to maintain. A Hadoop cluster is built using inexpensive commodity hardware and runs on traditional disk drives using a direct attached (DAS) configuration instead of an expensive storage area (SAN). The practice of storage redundancy makes the configuration more tolerable to hardware failures. Hadoop clusters are designed so that they are able to rebuild failed nodes easily [15].

Big unstructured data is more difficult to use. Traditional relational database users are able to access the data using a simple structured query language (SQL) that uses a sophisticated query engine that has been optimized to extract the data. Unstructured data is much more difficult to query. A sophisticated data user, such as a data scientist may be needed to manipulate the data. However tools are being developed to solve this problem. One tool is SparkSQL. This tool leverages conventional SQL for querying and works by converting SQL queries into MapReduce jobs. Another example is Microsoft Polybase which can join data from Hadoop and traditional databases and return a single result set [15].

To summarize, advances in Big Data technology, including data management of unstructured datasets and cloud computing are facilitating the development of platforms for more effectively capturing, storing, and manipulating large data sets sourced from multiple sources [29].

## 4.6 Big Data Trends for Healthcare

The costs for storing and parallel processing are decreasing [22]. Previously, we had to choose what data to capture and store because storage costs were so high. Now we can capture and store everything [17]. The use of the Internet of Things is growing. Internet connected technology is everywhere and has become a common and accepted part of our culture. For example, wearable fitness devices are continuously generating health information and sending it to the cloud.

Another trend is the establishment of standards and incentives in the industry that encourage the digitization and sharing of health care data. The Health Insurance Portability and Accountability Act (HIPAA) establishes national standards for electronic healthcare transactions for the submission of claims. Claims are the documents that health providers submit to insurance companies to get paid. Such standards encourage the widespread use of Electronic Document exchange. These standards have made it possible to effectively and easily share and exchange medical information between providers and insurance companies [22]. Medicare and Medicaid have set up Electronic Medical Record (EHR) incentive programs to encourage professionals and hospitals to adopt and demonstrate meaningful use of EHRs. The Affordable Health Care Act (ACA) encourages the shift from fee for service to value based payment structures by financing initiatives to test new payment models [33].

## 5 VALUE BASED REIMBURSEMENT

One of the most important strategies that we can take to reduce health care in the United States is to change the way that we reimburse providers from the traditional fee for service methodology to outcome based reimbursement. McKinsey and Company estimates that this strategy alone could reduce health care spending in the United States by 1 trillion dollars over the next decade [23]. This will also mitigate medical inflation because it will automatically promote preventative care and discourage the use of low value expensive technologies. Other benefits include: improved care coordination and the reduction of redundant care. All of this results in better health outcomes, and enhanced patient satisfaction.

With the fee for service payment structure providers are paid a fee for each and every service that they perform. This tends to

encourage overutilization instead of the efficient use of medical resources. The United States tends to perform more and more expensive diagnostic services and treatment services than any other country in the world. The United States is well known for over testing and over treatment [26]. Hospitals are rewarded for preventable readmissions. Physicians are rewarded as much for a failed medical procedure as they are for a successful one. It is up to each individual physician to determine what tests and treatment services to order. From a clinical perspective, many of these tests are not medically necessary. This is a wasteful use of resources.

The goal of value based reimbursement structures are to align payment incentives with the administration of efficient, high quality medical care. Basing provider reimbursement on performance and patient outcomes encourages providers to work towards optimizing patient health instead of just providing more health care services. Caregivers are also incentivized to be more innovative and to search for ways to improve health care delivery [5].

Many payers, including private health insurance companies, Medicare, and Medicaid are starting to base reimbursement on value based incentives. The Affordable Health Care Act includes provisions to encourage the development and adoption of more effective care delivery models. Some payers are also starting to reward pharmaceutical companies by basing reimbursements on drug effectiveness [18]. Systems that have been adopted to date include: patient centered medical homes, episode based payments, global payments, shared savings programs, value based contracting, and population models, including accountable care organizations.

In the patient centered home model, the primary care physician coordinates the patients care and is rewarded for improving quality and reducing costs for individual patients. Another value based system is a population model that rewards providers for improving the health of the entire population [20]. An example of this type of program is an Accountable Care Organization (ACO). In Accountable Care Organizations, groups of doctors, hospitals, and other providers work together to provide coordinated care for patients. In Medicare supported Accountable Care Organizations, providers share in Medicare savings when they deliver high quality care and manage costs wisely [7].

Big Data Analytics can play an integral role in the development and testing of new payment model methodologies. The development and adoption of such models are still in the infancy stage. Big Data Analytics has the potential to provide information that will result in innovative payment structure and reward insights. Big data can also play a role developing clinical best practices and in identifying reasons for unjustified clinical variability in current practices.

Big Data will help to support the implementation of models that have already been adopted. Value based health care depends upon quality data collection and precise data analytics [20]. First, the data must be collected and analyzed in order to define what defines quality care. Big Data is collected and analyzed in order to establish clinical guidelines that promote a more rational use of specific diagnostic tests and treatment protocols. Second, this information must be made available to health care givers in a format that they can use for day to day clinical decision making. This is often in the form of a cloud based integration platform [20]. Next, data must be collected on an ongoing basis to provide feedback indicating

whether the providers are meeting the defined standards and if not, what can be done to improve performance. In addition, the same data can benefit future patients when data analytics are taken beyond the initial reporting and are used to develop care protocols for entire patient populations [20].

One example is in which big data is being used to track and modify provider behavior is at Memorial Care, a six hospital system in Fountain Valley, California. Memorial Care uses physician performance analytics to analyze performance of hospital doctors and outpatient providers. So far, such tracking has resulted in the reduction 280 dollars per hospital stay for the average adult patient. This equates to a 13.8 million annual dollar savings for the Fountain Valley Hospital system [9].

## 6 ELECTRONIC HEALTH RECORDS

An Electronic Medical Record (EMR) is a digitized version of a patients medical chart. Whereas, an electronic medical record (EMR) typically includes information from one health provider, an electronic health record (EHR) includes information from multiple providers and documents all of the available information about the patient. The objective is to provide in one place, an electronic record of a patients health. This enables the sharing of information between providers. An electronic health record (EHR) contains medical history, diagnosis, medications, immunizations dates, allergy information, radiology images, and test results [36]. These records are made available to providers in real time. Electronic health record (EHR) systems often include electronic prescription subscribing systems. Also, they can include and be integrated with evidence based tools that help providers make immediate decisions about patients care. For example, an Electronic Health record system can also automatically check for problems such as medication conflicts and notify clinicians with alerts [13].

Electronic Health Records (EHRs) improve patient health care in so many ways. Physicians have better organized, more accessible, and more complete information about the patient. A clinicians ability to make an accurate diagnosis is improved. Easily accessible patient information reduces medical errors and unnecessary tests. There is a reduction in the incidence of duplicate tests. Coordination of care is improved because every caregiver is made aware of simultaneous care that is being provided by other caregivers. It easier to communicate critical clinical information to all applicable providers in a timely fashion. Because information is made available to providers in real time, there is a drastic reduction in the probability of errors caused by such things as allergic reactions or drug interactions, especially in emergency situations. Because electronic subscribing allows physicians to communicate directly with the pharmacies, prescriptions are no longer lost or misread [13]. Preventative care improves because it is easier to track and manage when patients are due for vaccinations and screenings. It becomes possible to track prescriptions to determine if a patient has been following doctors orders [34]. Productivity is increased, overlap care is reduced, and coordination of care is enhanced [5]. In general, electronic health records (EHRs) improve quality of care enhance patient safety, and contribute to better outcomes [13].

Electronic Health records (EHRs) have significantly improved the ability to treat chronically ill patients. In the past, providers

had to limit the decisions to the amount of information that was available to them at the time. The planning of care of a chronically diseased patient that had many symptoms was often mismanaged or delayed. Electronic health records (EHRs) enable the physicians to facilitate personalized treatment for these patients in a way that has never before been possible [5]. Providers have a comprehensive record of historical treatments, diagnostic data, medical history, and meticulous medical information all in one place [5]. The result is more efficient and effective treatment for chronically ill patients. There is a reduction in the number of potential side effects and an increase the patients quality of life all at a much reduced cost. [5].

Electronic health records (EHRs) also save money by reducing administrative costs. They reduce transcription costs and eliminate chart storage and access costs.

Between 2001 and 2014 Electronic Health record (EHR) usage in physician offices rose from 20 percent to 82 percent. According to Health Information Technology for Economic and Clinical Health (HITECH) research, electronic health records are being used in 94 percent of hospitals in the United States [34]. This amount of data that is being collected by large health systems and treatment centers around the country is massive [31].

## 7 PREDICTIVE ANALYTICS

### 7.1 Definition

Predictive analytics is the process of learning from historical data in order to make predictions about the future. The objective of predictive health analytics is to provide insights that enable personalized medical care for each individual patient [30]. Traditionally, physicians have always used predictive analytics, as they have always provided health care based upon what they know about the medical history of each individual patient. Predictive Health analytics seeks to supplement that knowledge with software tools that enable physicians to make more informed choices about the patients treatment based upon data from population cohorts [31]. Patients are directed to specific treatment plans based upon their specific conditions as compared to other patients in a similar cohort. This additional knowledge has the potential to provide physician with the information they need to provide a more effective treatment plans [31]. This becomes especially important for patients with complex medical histories who are suffering from multiple conditions [34]. Predictive analytics can also improve the accuracy of diagnosing patient conditions, better match treatments with outcomes, and better predict the specific patients at risk for disease [34].

Predictive analytics takes advantage of disparate data sources including: clinical, claims, research, sensors, social media, and genomic analysis.

Predictive analytics has the potential to materially reduce health care costs and improve patient care. Insights provided can in clinical decision support, prevent hospital readmission preventions, aid in adverse incidence avoidance, and help chronic disease management. In addition, predictive analytics can identify treatments and programs that do not deliver demonstrable benefits or that cost too much [29]. Some predictive models reduce readmissions by identifying environmental of lifestyle factors that increase risk

or trigger adverse events so that treatment plans can be adjusted according. [29].

## 7.2 Patient Profile Analytics

Patient Profile Analytics is a specific type of predictive analysis in which patient profiles are developed to identify individuals who may be at risk for developing a disease and who could benefit from proactive management, such as lifestyle modifications. For example, patient profile analytics can be used to identify patients who may be at risk for developing diabetes.

## 7.3 Risk Stratification

One area in which predicting patients at risk can yield the greatest results is in identifying the patients who are at the greatest risk for the most adverse outcomes or costliest diseases [29]. Risk stratification is a methodology that can be used to identify and track the sickest and potentially costliest patients. The tool ranks or stratifies patients by potential risk and flags high risk cases for additional management. A risk stratification predictive tool takes into account risk factors such as missed doctors appointments in addition the symptoms. The tool enables doctors to intervene earlier to avoid hospital admissions and costly treatment [9].

## 7.4 Predictive Analytic Examples

Hundreds of thousands of dollars are spent on cancer care. Big data can be used to develop individualized, personalized cancer care programs. There is a web based application, which was sponsored by the National Cancer Institute that uses data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening trial together with patient risk factor and demographic data to help develop patient specific treatment regimens [6].

Congestive heart failure accounts for more medical spending than any other diagnosis. The earlier this condition is diagnosed, the easier it is to treat and to avoid dangerous and expensive complications. However, early manifestation is difficult to recognize and can easily be missed by physicians [22]. Machine learning algorithms have the ability to take into account many more factors than doctors alone. Predictive modeling and machine learning using large sample sizes can identify nuances and patterns that were previously impossible to see. As a result, machine learning models in the form of predictive analytics substantially improved clinicians ability to accurately diagnose persons with congestive heart failure [34].

Optum labs has developed a database with the electronic health records of over 30 million patients. They use the database to develop predictive analytic tools, the objective of which is to help doctors make Big data informed decisions that will improve patients treatment [22].

Parkland Hospital in Dallas, Texas uses predictive modeling to identify high risk patients in the coronary care unit and to predict likely outcomes when the patients are sent home. To date, Parkland has reduced readmissions for Medicare patients with heart failure by 31 percent. This equates to a 500000 dollar annual savings for this one hospital [9].

## 8 INTERNET CONNECTED MEDICAL DEVICES

Internet connected medical devices are becoming more affordable and are being used more and more commonly. Gartner, the analysis firm, estimates that there will be more than 25 billion connected health devices by the year 2020 [15]. These devices collect data in real time and send information into the cloud. Devices include blood pressure monitors, pulse oximeters, glucose monitors, and electronic scales [15]. Some of these devices are being used as preventive care devices. Other devices are being used by health care providers to aid in the monitoring of patient conditions. Big Data is required because the process involves the capture and analysis of large volumes of fast moving data from in hospital and in home devices in real time.

### 8.1 Preventative Care

Millions of people are using mobile technology help live healthier lifestyles. Smart phone applications together with wearable devices such as Fitbit, Jawbone, and Samsung Gear Fit are designed to track the wearers exercise and activity levels [12]. Measures that are typically tracked include: the number of steps taken, number of calories burned, and number of stairs climbed. The objective is to encourage the users to take a more active role in their own health and wellbeing by being more physically active. Such devices can provide individuals with the information that they need to make more informed decisions, better manage their health, and to more easily track and adopt healthier behaviors [3]. In the future, it is conceivable that it will be routine to share this information with personal physicians and that it will be incorporated into regular health care management.

An individuals data can be uploaded from the device to the cloud where it is aggregated with information from other users [15]. In an initiative between Apple and IBM, a big data platform is being developed that will allow iPhone and Apple Watch users to share their data with IBMs Watson Health cloud health care analytics service. The information will use the combination of real time activity information in combination with biometric data to discover new medical insights [12].

### 8.2 Medical Monitoring

Remote monitoring enable medical professional to monitor a patient remotely using various technological devices. The devices can be worn by patients with health conditions at home and in medical facilities to stream data continuously to provide real time remote patient monitoring. The devices can improve care by giving patients the ability to self-manage their conditions. Processing of real time events can be supplemented with machine learning algorithms to help provide physicians with information they need to make lifesaving interventions [22]. Patient care tends to be more proactive as patient vital signs are can be monitored constantly [22]. Medical alerts can be sent to care providers such that they immediately aware of changes in a patients condition and can respond accordingly. Devices are often used for adverse risk prediction. Remote monitoring is typically used to monitor conditions such as heart disease, diabetes mellitus, and asthma. One example of the

use of personal devices in patient care is pediatricians monitoring asthmatics to identify environmental triggers for attacks [6].

Real time systems analysis improves patient care while simultaneously reducing health care costs [5]. The devices are especially advantageous to individuals who reside in remote areas. Other advantages include: a reduced incidence of severe events, improved in patient safety, and high patient satisfaction levels.

## 9 PUBLIC HEALTH

Data science is being used in cities throughout the United States to predict and impede potential public health issues before they even start. For example, the Chicago Department of Public Health is modeling a program to target lead exposure in children. Information is collected from multiple sources such as, home inspection records, assessor values, health records, and census data. Predictive analytic algorithms then determine which houses have the highest potential risk. This information is then being incorporated into Electronic health records (EHRs) to automatically alert physicians to possible lead exposure risk concerning their pediatric and pregnant patients. Chicago has similar programs in place for food protection and tobacco control [14].

In San Diego, California the public health department routinely gathers big data health related information and publishes it on a user friendly web site. Information is gathered from sources such as marketing companies, mobile apps and demographic data. The data includes everything from vegetable consumption to diabetes occurrences. In one initiative, Live Well, the information was able to reduce the obesity rates at a local elementary school by 5 percent. A project that is currently in progress is the study and analysis of areas that have high rates of Alzheimers [19].

## 10 TRANSPARENCY

In the United States, health care price information is rarely made available to the health care consumers when they receive the care. Patients usually become aware of the costs when they receive the bill. The price of health procedures can vary radically by provider. Prices can even vary by payer for the same provider. In one study, it was estimated that consumers paid 10 to 17 percent less when they were given access to comparative price data. According a paper that was published by the American Economic Journal Economic Policy, if patients had access to price data and were willing to shop around, they could be pay significantly less for everything from routine screenings to knee surgery [2]. This tended to work best for consumers who had to pay for at least some portion of their own care.

Online pricing is a potential Big Data solution. Health related price web sites provide approximate prices for health services and procedures in fairly transparent formats. Online resources are now being made available by insurers, government agencies, internet companies and medical care providers. National insurers such as Anthem, United Health group, Humana, Aetna, and Cigna offer pricing tools to their customers. Some states, including New Hampshire, Maine, Oregon, and Massachusetts publish health pricing websites. The internet company Healthcarebluebook.com publishes information for all consumers in the United States [35].

The trend towards pay for performance reimbursement agreements will also help the cost transparency issue. This is because these pricing structures encourage health care providers to share information [5].

## 11 EVIDENCE BASED MEDICINE

Evidence based medicine (EBM) is an approach to medical practice that emphasizes the use of evidence from well designed and well conducted research to optimize decision making [37]. Evidence based medicine is an approach that supplements a clinicians knowledge, which may be limited by knowledge gaps or bias, with the formal and explicit information such as scientific literature or best practice methodology. Evidence based medicine eliminates guesswork for health care providers. Instead of having to rely only on their own personal judgement, providers can base treatment and protocols on credible scientific data [5].

Big Data analytics supports the research and development of evidence based best practice treatment protocols. Structured and unstructured data from a variety of sources is combined and big data algorithms are applied. Sources may include electronic medical records, financial and operational data, clinical data, and genomic data [29]. The aggregating individual data sets into big data sets enable analysis for conditions that typically have small populations. An example is the study of individuals with gluten allergies [18].

## 12 DRUG COSTS

It is a well known fact that drugs in the United States are priced higher than they are in other countries. There are many complicated contributing factors. One factor is lack of price regulation. Another factor is the economic structure of the health care system. Because the system includes multiple payers, there is no one payer with the power to effectively negotiate with the pharmaceutical companies as there are in other economies. Therefore, drug companies typically set drug prices at whatever the market will bear. Newly developed drugs usually have higher price tags. Big Data analytics cannot fix all of the problems with the drug market, but there are some areas in which it may have an impact: medication therapy management capabilities, drug comparison technology, and pharmaceutical research and development process improvements [4].

### 12.1 Medication Therapy Management

Big data analytics can play a significant role in improving the Medication Therapy Management process. Adverse drug events cost billions of dollars and result in thousands of patient deaths. Physicians and pharmacist are often overwhelmed to the point of not having the time to implement appropriate drug therapies. Drug therapies are becoming more difficult to manage as more patients are taking multiple medications. Big Data cloud analytics are helping clinicians better co manage drug therapies, and to identify drug interactions, adverse side effects, and additive toxicities in real time. The results include a reduction in the number of patient deaths, emergency room visits, hospital admissions, and hospital readmissions [9].

## **12.2 Comparison of Competitor Drugs**

In the research, there tends to be a lot of information about individual drugs. However, there is not much information about how drugs perform in comparison to their competitors. There needs to be more drug comparative information so that physicians are better informed about the true benefits of prescribing a more costly medication as compared to a less expensive or generic drug [4]. Big data technology can play a role in making such comparisons easier to accomplish.

## **12.3 Pharmaceutical Research and Development**

Big Data can help to streamline the Pharmaceutical Research and development process. As a result, important drugs can be delivered to the market more quickly and the cost of drug development will be reduced.

Big data can enhance the process of identifying appropriate patients to enroll in the clinical trials. First, multiple sources are now available from which to select patients. For example, social media can be incorporated into the selection process and used in addition to physician information. Secondly, the participate selection criteria can include more inclusive factors, such as genetic information. This will enable better targeting of potential trial subjects which will result in more pertinent information, while at the same time shorting trail times and reducing expenses [24].

Trial can be monitored and tracked in real time. Real time trial monitoring can decrease the number of safety and operational issues. The result is the avoidance of potentially costly issues such as adverse events or unnecessary delays [24].

Electronically captured data can improve communication. Information can be shared easily between functions and external parties. All interested individuals can have access to the data at the same time including all departments, external partners, physicians, and contract research organizations (CROs). This will replace the issue of having rigid departmental data silos that hinder interaction [24].

Genomic and proteomic data can be used to speed drug development by providing the capability to better target treatments based upon genetic indicators [17].

## **13 ADMINISTRATIVE COSTS**

According to the Institute of Medicine (IOM), the United States spends 361 billion annually on health care administration. This is more than twice our total spending on heart disease and three times our spending on cancer. Also according to the IOM, fully half of these expenditures are unnecessary [9].

One way that providers can save money is to digitize billing processes such as benefit verification, denial management, and claims submission. A benefit verification that is done electronically costs 49 cents per patient. Comparatively, the same process done manually costs 8 dollars. It is estimated that providers could save 9.4 dollars annually by transitioning to electronic processing [21].

One example in which digitized processes are being used to streamline billing processes effectively is at the Phoenix Childrens Hospital in Arizona. They use a tool that automatically converts the clinical notes in the electronic health record (EHR) system to billable diagnostic codes [21].

## **14 FRAUD AND ABUSE**

Common types of fraud and abuse include: billing for services that are not rendered, billing for more expensive procedures than were actually delivered, and the performance of unnecessary services.

In the past, the process of identifying misrepresented claims was tedious and time consuming. Big Data analytics makes it possible to easily identify and tag such claims. According to an article by RevCycle Intelligence, when there is repeated misrepresentation of some key fact or event, patterns are created in the data that can be detected by comparing the information to legitimate claims [10]. Anthem Health Insurance, one of the nations biggest insurance payers, uses big data and machine learning algorithms to tag suspicious claims as the claims are being processed. Tagged claims are then sent to clinical coding experts for review. The objective is to identify and address fraudulent claims before they are actually paid [10].

The Center for Medicare and Medicaid Services used predictive data analytics to identify and recover 210.7 million [22] in health care fraud in 2015. They did this by assigning risk scores to claims and providers via algorithms. This enabled the identification of abnormal billing patterns in claim submissions [10].

United Healthcare realized a 2200 percent return on their investment in a Hadoop Big Data platform that was used to identify and tag inaccurate claims using a systemic and repeatable methodology [22].

Other uses of Big Data analytics in fighting fraud and abuse include: identifying links between providers to access whether an identified unethical activity is being practiced by related providers, identification of a hospitals overutilization of services in a short time period, recognizing patients who are receiving health care services from different hospitals in different locations at the same time, and detecting prescriptions that are filled for the same patient in multiple locations at the same time. Big Data analytics can also utilize machine learning algorithms combined with historical information to detect trends in anomalies and suspicious data patterns.

## **15 GENOMICS ANALYTICS**

Big data is playing a major role in the field of genomics and precision medicine. These technologies are helping clinicians choose the best treatment plan for individuals based upon their genetic makeup. Combining data from electronic health records (EHRs), clinical trials, and genetic testing gives researchers information to develop more effective treatments for complex diseases such as cancer and diabetes [25], and HIV. Genetic testing that has been made possible by the mapping of the human genome will cut costs and improve survival rates [1].

One area in which genomics can have a dramatic impacts is in pharmaceuticals management. In the United States, 300 million dollars are spent annually on pharmaceuticals. Studies indicate that between 20 to 75 percent of patients are not responsive to prescribed drug therapies. This can often be contributed to incorrect dosing or drug mismatches. However, 50 percent of the time it is because of a molecular mismatch between the patient and the drug. According to Alan Mertz, president of the American Clinical Laboratory Association, an estimated 30 to 110 billion can be saved

by using genetic test to select a drug that is a precise match for the genetics of the patient. By using each patients unique genomic profile, therapy can become more targeted and the instances of inappropriate care will be reduced [1].

For breast cancer patients, genetic testing can identify which 30 percent of women of an overabundance of the HER2 protein. Regular chemotherapy will not help these women, but a drug called Herceptin does. Having this information not only provides doctors with the information they need to prescribe the correct medication, it enables thousands of women avoid needless harsh, expensive chemotherapy treatment. As a result, genetic testing has been shown to reduce the risk of death by 33 percent and the risk of recurrence by 52 percent for breast cancer patients. The resulting savings are estimated to be 24 thousand dollars per patient [1].

Genetic tests can help physicians select the appropriate drug for patients with metastatic colon cancer. According to one estimate, 700 million dollars could be saved annually be obtaining this information before administering treatment [1].

According to a 2006 Brookings/AEI estimate, using genetic tests to determine the appropriate dose of the blood thinner, warfarin, could save the United States 1.1 billion dollars annually. According to a study in June 2010 by the Journal of American College of Cardiology, this test could reduce hospital admissions that are caused by inaccurate dosages by 31 percent [1].

Genomic technology is also good for the United States economy. According to Battelle, a global research organization, human genome sequencing projects generated 796 billion in economic output, 244 billion in personal income and 3.8 million job-years of employment in the United States [1].

The process of gene sequencing continues becomes more efficient and cost effective. It is expected to become a regular part of medical care in the near future [15].

## 16 TELEMEDICINE

Telemedicine is receiving medical treatment and advice remotely, on a computer over the internet with a physician [12]. Telemedicine has been in the market for 40 years, but the with availability of internet connected technology such as smartphones, wireless devices, and video conferences, it is becoming commonplace. It is primarily used for initial diagnosis, remote patient monitoring, and medical education. However, it is also being used for more complicated care such as telesurgery. Telesurgery is a technique in which doctors perform surgery via robots with the assistance of high speed real time data delivery technology [34].

Telemedicine is especially beneficial to patients who live in rural communities who may have to travel long distances to see a doctor or specialist. Telemedicine also gives doctors who are located in multiple locations the ability to discuss and share information. Telemedicine facilitates medical education by giving caregivers the ability to observe and be trained by subject experts no matter where their location.

Telemedicine has the potential to significantly reduce costs by reducing the number of outpatient and hospital visits [38].

## 17 USE CASES

Valence Health has built a data lake that they use as their primary data repository using a MapR Converged Data Platform. The system includes 3000 inbound data feeds and contains 45 different types of data including: lab test results, patient vitals, prescriptions, immunizations, pharmacy benefits, claims information from doctors and hospitals. The system reports dramatically better system performance than legacy system technology. For example, previously, it took 22 hours to process 20 million laboratory records. Now the processing time for the same number of records is 20 minutes. In addition, the new system requires less hardware [22].

The National Institute of Health developed a data lake which combines data sets from separate institutions. Now that all of the data is housed in the same location, analysis is more efficient and can be more easily shared [22].

United Healthcare uses Hadoop to maintain a platform with tools that they use to analyze information generated from claims, prescriptions, provider contracts, plan subscriber, and review information [22].

Novartis, a global healthcare company, uses Hadoop and Apache Spark to build a workflow system that aids in the integration, processing, and analysis of Next Generation Sequencing research as it relates to Genomic Analytics [22].

## 18 CHALLENGES

One of the most compelling challenges is clinicians willingness and ability to change behavior based upon the information provided by the data. Studies have shown that it takes more than a decade of compelling clinical evidence before a new finding becomes common clinical practice. Therefore, we need to do a better job of working with clinicians on finding ways to use the data to provide higher quality care [17].

In health care, the privacy, security, and confidentiality of the patient is paramount [15]. Big data technology has inconsistent security technology. The Health Insurance Portability and Accountability Act (HIPPA) is a federal law that was passed in 1996 that sets a national standards to protect the confidentiality of medical records and personal health information. The HIPAA law is applicable to any component of the information can be used to identify a person. The protections apply to both electronic and non-electronic forms of information [32]. HIPAA regulations make it a federal offense to breach patient security. It is important to work with vendors who understand the importance of security [15]. Liason Technologies is one company that provides solutions to the healthcare and life sciences industry that has experience meeting the HIPAA security requirements [22].

Health care data has inconsistent formatting and definitional issues [17]. There is proliferation of data formats and data representations. There are inconsistent variable definitions. A value may have different meanings for different groups. For example, a cohort definition for an asthmatic patient often differs from one group of clinicians to another [16]. Big data has the challenge of bringing all of this information together.

Another issue is lack of technical experts. The manipulation and extraction of data from often unstructured data sets require special knowledge. There have been some recent changes in tooling that

will make it easier for individualized with less specialized skills to manipulate the data. For example, Big data is starting to use include SQL as a tools for querying and data manipulation. Examples are Microsoft Polybase, Impala, and SQL Hadoop [15].

## 19 CONCLUSION

Big data analytics has huge potential to save the United States billions of dollars in health care costs while drastically improving health outcomes. Vast amounts of information is being captured, stored and combined in ways that offer insights have never before been possible. Innovative Big data tools are reducing medical waste, decreasing medical errors, fighting fraud, and keeping people healthier. Value based reimbursement solutions have the potential to revolutionize the health delivery system in the United States by motivating providers to find ways to deliver the best possible medical care with the most economical use of resources. The development of most of these tools is only in the preliminary stage. Therefore, we are only beginning to realize some of the potential benefits. Big data really does have the potential to bend the cost curve. Big data in health care is here to stay.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the teaching assistants in the Data Science department at Indiana University for their support and suggestions to write this paper.

## REFERENCES

- [1] American Clinical Library Association. 2011. Genetic Testing Can Help the United States Cut Costs and Improve Care. Web page as article. (July 2011). <https://www.prnewswire.com/news-releases/genetic-testing-can-help-the-us-cut-costs-and-improve-health-care-126105103.html>
- [2] American Economic Association. 2017. Would Price Transparency Lower Health-care Costs. Web page as article. (Feb. 2017). <https://www.aeaweb.org/research/health-care-price-transparency>
- [3] Tanya Bentley. 2018. Waste in the US Health System - A conceptual framework. *The Milbank Quarterly* 86 (Dec. 2018), 629–659. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690367/>
- [4] Business Insider. 2016. Why the Price of Prescription drugs in the US is Out of Control. Web page as article. (Aug. 2016). <http://www.businessinsider.com/why-the-us-pays-more-for-prescription-drugs-2016-8>
- [5] Christian Ofori Boateng. 2016. Top 3 Ways Big Data Helps Decrease the Cost of Health Care. Web page as Article. (Nov. 2016). <https://go.christiansteven.com/top-3-ways-big-data-helps-decrease-the-cost-of-health-care>
- [6] CIO. 2015. How Big Data can save 400 billion in healthcare costs. Web page as Article. (Oct 2015). <https://www.cio.com/article/2993986/big-data/how-big-data-can-help-save-400-billion-in-healthcare-costs.html>
- [7] CMS Centers for Medicare and Medicaid Services. 2017. Accountable Care Organizations. Web page. (Nov. 2017). <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/ACO/>
- [8] Consumer Reports. 2014. Why is Healthcare so Expensive. Web page. (Sept. 2014). <https://www.consumerreports.org/cro/magazine/2014/11/it-is-time-to-get-mad-about-the-outrageous-cost-of-health-care/index.htm>
- [9] DataFloq. 2016. Five ways Big Data in reducing healthcare costs. Web page as article. (March 2016). <https://datafloq.com/read/5-ways-big-data-reducing-healthcare-costs/89>
- [10] Datameer. 2017. The Role of Big Data in Preventing Healthcare Fraud, Waste, and Abuse. Web page as article. (Sept. 2017). <https://www.datameer.com/company/datameer-blog/role-big-data-preventing-healthcare-fraud-waste-abuse/>
- [11] Digitalist. 2016. Can Big Data Analytics Save Billions in Healthcare Costs. Web page as Article. (Feb. 2016). <http://www.digitalistmag.com/resource-optimization/2016/02/29/big-data-analytics-save-billions-in-healthcare-costs-04037289>
- [12] Forbes. 2015. How Big Data in changing Healthcare. Web page as Article. (April 2015). <https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#427274d12873>
- [13] Forbes. 2016. How an Electronic Health Record can Save Time, Money and Lifes. Web page. (Dec. 2016). <https://www.forbes.com/sites/robertpearl/2016/12/01/how-an-electronic-health-record-can-save-time-money-and-lives/2/#4445b8275f57>
- [14] Harvard Business Review. 2014. How Cities are Using Analytics to Improve Public Health. Web page as article. (Sept. 2014). <https://hbr.org/2014/09/how-cities-are-using-analytics-to-improve-public-health>
- [15] Health Catalyst. 2017. Big Data in Healthcare Made Simple: Where it Stands Today and Where its Going. Web page as Article. (Oct. 2017). <https://www.healthcatalyst.com/big-data-in-healthcare-made-simple>
- [16] Health Catalyst. 2017. Five Reasons Healthcare Data is so Complex. Web page as article. (Nov. 2017). <https://www.healthcatalyst.com/>
- [17] Health Catalyst. 2017. Hadoop in Healthcare A no nonsense Q and A. Web page as article. (Nov. 2017). <https://www.healthcatalyst.com/Hadoop-in-healthcare>
- [18] Kayyali, Basel, Knott, David, Kuiken, Steve Van. 2013. McKinsey on Healthcare. Web page as Article. (April 2013). <http://healthcare.mckinsey.com/big-data-revolution-us-healthcare>
- [19] KQED Science. 2015. How San Diego is Using Big Data to Improve Public Health. Web page as article. (Aug. 2015). <https://ww2.kqed.org/futureofyou/2015/08/19/how-san-diego-is-using-big-data-to-improve-public-health/>
- [20] Liason. 2017. Value Based Healthcare - The patient is the Center but Data is the Key. Web page as blog. (June 2017). <https://www.liason.com/blog/2017/06/22/value-based-healthcare-patient-center-data-key/>
- [21] Managed Healthcare Executive. 2017. Five ways to reduce healthcare administrative costs. Web page as article. (April 2017). <http://managedhealthcareexecutive.modernmedicine.com/managed-healthcare-executive/news/five-ways-reduce-healthcare-administrative-costs>
- [22] McDonald, Carol. 2016. How Big Data is Reducing Costs and Improving Outcomes in Healthcare. Web page as Article. (June 2016). <https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/>
- [23] McKinsey and Company. 2013. The Trillion Dollar Prize. Web page as article. (Feb. 2013). <https://healthcare.mckinsey.com/sites/default/files/the-trillion-dollar-prize.pdf>
- [24] McKinsey and Company. 2017. How Big Data can Revolutionize pharmaceutical R and D. Web page as article. (Nov. 2017). <https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d>
- [25] Pacient. 2017. How Big Data Can Improve Health Care. Web page as article. (Nov. 2017). <https://pacient.care/decks/privacy-technology/health-technology/how-big-data-can-improve-healthcare>
- [26] PBSO News Hour. 2012. Health Costs: How the US Compares with Other Countries. Web page as Article. (Oct. 2012). <https://www.pbs.org/newshour/health/health-costs-how-the-us-compares-with-other-countries>
- [27] Practice Fusion. 2017. EHR Adoption Rates 20 Must see stats. Web page as Article. (March 2017). <https://www.practicefusion.com/blog/ehr-adoption-rates/>
- [28] OECD Publishing. 2017. *Health at a Glance 2017*. OECD, Paris. [http://dx.doi.org/10.1787/health\\_glance-2017-en](http://dx.doi.org/10.1787/health_glance-2017-en)
- [29] Raghupathi Viju Raghupathi, Wullianallur. 2014. Big Data Analytics in Healthcare Promise and Potential. *Springer Health Information Science and Systems* 2 (Feb. 2014), 2–3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/>
- [30] Rock Health. 2017. The Future of Personalized Healthcare: Predictive Analytics. Web page. (Nov. 2017). <https://rockhealth.com/reports/predictive-analytics/>
- [31] Search Technologies. 2017. Using Big Data Predictive Analytics to Improve Healthcare. Web page as article. (Sept. 2017). <https://www.searchtechnologies.com/blog/predictive-analytics-in-healthcare>
- [32] Stephen B Thacker. 2003. HIPAA Privacy Rule and Public Health. *CDC* 52 (April 2003), 1–12. <https://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm>
- [33] The Common Wealth Fund. 2017. The Affordable Care Acts Payment and Delivery System reforms: A progress report. Web page as article. (Feb. 2017). <http://www.commonwealthfund.org/publications/issue-briefs/2015/may/aca-payment-and-delivery-system-reforms-at-5-years>
- [34] The datapine blog. 2017. Nine examples of Big Data Analytics in Healthcare that Can Save People. Web page. (May 2017). <https://www.datapine.com/blog/big-data-examples-in-healthcare/>
- [35] The Wall Street Journal. 2017. How to Research Medical Prices. Web page as article. (Nov. 2017). <http://guides.wsj.com/health/health-costs/how-to-research-health-care-prices/>
- [36] US Department of Health and Human Resources. 2017. EHR Basics. Web page. (Sept. 2017). <https://www.healthit.gov/providers-professionals/learn-ehr-basics>
- [37] Wikipedia. 2017. Evidence Based Medicine. Web page. (Nov. 2017). [https://en.wikipedia.org/wiki/Evidence-based\\_medicine](https://en.wikipedia.org/wiki/Evidence-based_medicine)
- [38] Wikipedia. 2017. Telemedicine. Web page. (Nov. 2017). <https://en.wikipedia.org/wiki/Telemedicine>
- [39] Zane Benefits. 2017. FAQ - How much does Individual Insurance cost. Web page. (Nov. 2017). <https://www.zanebenefits.com/blog/bid/97380/faq-how-much-does-individual-health-insurance-cost>

```
bibtext report
```

```
=====
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-12-16 09.38.25] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Typesetting of "report.tex" completed in 1.1s.
./README.yml
53:20     error      no new line character at the end of file  (new-line-at-end-of-file)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Judy Phillips
hid: 332
paper1: Oct 31 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

```
-----
```

wordcount

---

10  
wc 332 project 10 8955 report.tex  
wc 332 project 10 9312 report.pdf  
wc 332 project 10 1543 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

6: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0

True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not cahnge the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
The following tests are optional
=====
```

Tip: newlines can often be replaced just by an empty line

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics

Sean M. Shiverick

Indiana University Bloomington

smshiver@indiana.edu

## ABSTRACT

Classification of opioid misuse and abuse can identify important features relevant for predicting drug addiction and overdose death. Machine learning procedures were applied to data from a large National Survey of Drug Use and Health (NSDUH-2015) to classify individuals for illicit opioid use according to demographic characteristics and mental health attributes (e.g., depression). Classification models of opioid addiction can be extended for big data health analytics to include high-dimensional datasets, data collected over previous years, or expanded to the larger population of patients taking prescription opioid medication. The results seek to raise awareness of risk factors related to opioid addiction among patients and medication prescribers, and help decrease the risk of opioid overdose death.

## KEYWORDS

Big Data, Health Analytics, Classifier Algorithms, Opioid Addiction, i523, hid335

## 1 INTRODUCTION

Big Data offers tremendous potential to fuel innovation and transform society. Can this momentum be harnessed to address a serious health crisis such as the opioid overdose epidemic? [7] Health informatics is generating huge amounts of data at a rapid pace, from electronic medical records (EMRs), clinical research data, to population-level public health data [5]. This project considers health analytics from two levels, the research questions being addressed and the data used to answer them. The question of interest in this project is whether opioid dependency and addiction can be predicted from demographic attributes and psychological characteristics. Survey research provides data on a wide range of issues that people may be reluctant to disclose, including mental health disorders, personal medical health concerns, prescription medications, and illicit drug use. Responses to surveys may be biased to some degree, but measures of confidentiality and anonymity help to assure more accurate disclosures. The goal of this project is to use machine learning procedures to classify individuals susceptible to opioid abuse and dependence. Understanding the features that contribute to opioid addiction can identify underlying risk factors and increase awareness of potential opioid abuse for patients and health care providers. The results could be extended to big data from previous years of the opioid crisis and to the larger population of patients taking prescription opioid mediation. Different machine learning classification methods are discussed.

## 1.1 Opioid Overdose Epidemic

The abuse of prescription opioid medication in the U.S. has become a major health crisis of epidemic proportions [26]. Over 2 million Americans were dependent or abused prescription opioids such as oxycodone or hydrocodone in 2014[3]. Overdose deaths from prescription opioids have quadrupled since 1999, resulting in more than 180,000 deaths between 1999 to 2015 [11]. Drug overdose deaths increased significantly for males and females, between 25-44 years, ages 55 and older, for Non-Hispanic Whites and Blacks, in the Northeast, Midwest, and Southern regions of the U.S. [7]. Mobile health applications can monitor patient medication consumption and provide an early warning system for potential abuse, detecting sudden changes in medications, higher dosages, or rapid escalation of a prescribed dosage [25]. Reliable information about medication dosages can be difficult to obtain based on self-reports. Individuals dependent or addicted to prescription opioids may obtain synthetic opioids such as fentanyl or illicit drugs such as heroin. Because the dosage levels and potency of illicit opioids are largely unknown, there is greater risk of drug overdose death. The sharp increase in overdose deaths due to synthetic opioids (other than methadone) has coincided with the increased availability of illicitly manufactured fentanyl, which is indistinguishable from prescription fentanyl. The findings indicate the opioid overdose epidemic is getting worse, and requires urgent action to prevent opioid dependence, abuse and overdose death. The target group for this project is individuals who reported misusing or abusing prescribed opioid medication who also used heroin, shown in Figure 1.

## 1.2 Machine Learning Approaches

Machine learning is a set of procedures and automated processes for extracting knowledge from data. The two main branches of machine learning are supervised learning and unsupervised learning. Supervised learning problems involve prediction about a specific target variable or outcome of interest. If a given dataset has no target outcome, unsupervised learning methods can be used to discover underlying structure in unlabeled data. The goal of this project is to classify opioid addiction and focuses on supervised learning. Supervised learning is used to predict a certain outcome from a given input, when examples of input/output pairs are available [10]. A machine learning model is constructed from the training set of input-output pairs, to predict new test data not previously seen by the model. The two major approaches to supervised learning problems are regression and classification. When the target variable to be predicted is continuous, or there is continuity between the outcome (e.g., home values, or income), a regression model is used to test the set of features that predict the target variable. If the target is a class label, set of categorical or binary outcomes (e.g., spam or ham, benign or malignant), then classification is used to

predict which class or category label that new instances will be assigned to.

### 1.3 Classification Algorithms

Comparing the performance of different learning algorithms can be helpful for selecting the best model for a given problem [14]. One of the simplest classification algorithms is K-Nearest-Neighbors (KNN) which takes a set of data points and classifies a new data point based on the distance (e.g., Euclidean, by default) to its nearest neighbors. The main parameter for KNN is the number of neighbors, and k of 3 or 5 neighbors works well. The advantage of the KNN classifier is that it provides a solution that is easy to understand. A limitation of KNN is that it does not perform well with a large number of features (100 or more) or sparse datasets. Several different classification algorithms are considered below.

**1.3.1 Logistic Regression Classifier.** Logistic regression is a commonly used linear model for classification problems. The decision boundary for the logistic regression classifier is a linear function of the input; a binary classifier separates two classes using along a line, plane, or hyperplane. Linear classification models differ in terms of (1) how they measure how well a particular combination of coefficients and intercept fit the training data, and (2) the type of regularization used [10]. The main parameter for linear classification models is the regularization parameter C. High values of C correspond to less regularization and the model will fit the training set as best as possible, stressing the importance of each individual data point to be classified correctly. By contrast, with low values of C, the model puts more emphasis on finding coefficient vectors (i.e., weights) that are close to zero, trying to adjust to the majority of data points. In addition, the penalty parameter influences the coefficient values of the linear model. The L2 penalty (Ridge) uses all available features, but pushes the coefficient values toward zero. The L1 penalty (Lasso) sets the coefficient values for most features to zero, and uses only a subset of features for improved interpretability. This analysis used a logistic regression classifier to predict Heroin use from demographic attributes, mental health, prescription opioids, medication use, misuse, and illicit drug use.

**1.3.2 Tree Based Models.** Decision tree models are widely used for classification and regression. Tree models “learn” a hierarchy of if-else questions that are represented in the form of a decision tree. Building decision trees proceeds from a root node as the starting point and continues through a series of decisions or choices. Each node in the tree either represents either a question or a terminal node (i.e., leaf) that contains the outcome. Applied to a binary classification task, the decision tree algorithm *learns* the sequence of if-else questions that arrives at the outcome most quickly. For data with continuous features, the decisions are expressed in the form of, “Is feature x larger than value y?” [10] In constructing the tree, the algorithm searches through all possible decisions or tests, and find a solution that is most informative about the target outcome. A decision tree classifier is used for binary or categorical targets, and decision tree regression is used for continuous target outcomes. The recursive branching process of tree based models yields a binary tree of decisions, with each node representing a test that considers a single feature. This process of recursive partitioning

is repeated until each leaf in the decision tree contains only a single target. Prediction for a new data point proceeds by checking which region of the partition the point falls in, and predicting the majority in that feature space. The main advantage of tree based models is that they require little adjustment and are easy to interpret. A drawback is that they can lead to very complex models that are highly overfit to the training data. A common strategy to prevent overfitting is *pre-pruning*, which stops tree construction early by limiting the maximum depth of the tree, or the maximum number of leaves. One can also set the minimum number of points in a node required for splitting. Another approach is to build the tree and then remove or collapse nodes with little information, which is called *post-pruning*. Decision trees work well with features measured on very different scales, or with data that has a mix of binary and continuous features.

**1.3.3 Random Forests Classifier.** A random forest is a collection of decision trees that are slightly different from the others, which each overfits the data in different ways. The idea behind random forests is that overfitting can be reduced by building many trees and averaging their results. This approach retains the predictive power of trees while reducing overfitting. Randomness is introduced into the tree building process in two ways: (a) selecting a bootstrap sample of the data, and (b) selecting features in each node branch [10, 14]. In building the random forest, we first decide how many trees to build (e.g., 10 or 100), and the algorithm makes different random choices so that each tree is distinct. The bootstrapping method repeatedly draws random samples of size n from the dataset (with replacement). The decision trees are built on these random samples that are the same size as the original data, with some points missing and some data points repeated. The algorithm also selects a random subset of p features, repeated separately each node in the tree, so that each decision at the node branch is made using a different subset of features. These two processes help ensure that all of the decision trees in the random forest are different. The important parameters for the random forests algorithm are the number of sampled data points and the maximum number of features; the algorithm could look at all of the features in the dataset or a limited number. A high value for *maximum-features* will produce trees in the random forest that are very similar and will fit the data easily based on the most distinctive features, whereas a low value will produce trees that are very different from each other, and reduces overfitting. Random forests is of the most widely used ML algorithms that works well without very much parameter tuning or scaling of data. A limitation of this approach is that Random forests do not perform well with very high-dimensional, data that is sparse data, such as text data.

### 1.4 Project Goals

The general idea of the project is that prescription opioid dependency and addiction will in many cases lead to the use of illicit opioids such as heroin or fentanyl. According to this reasoning, it was hypothesized that individuals who report using heroin may also be susceptible to misusing or abusing prescription opioid medications. The goal of the study was to identify the set of features important for predicting opioid addiction. The data used in the project is from the National Survey on Drug Use and Health from 2015 (NSUH-

2015) [1], which is the most recent year available. The NSDUH-2015 is a comprehensive survey that covers all aspects of substance use, misuse, dependency, and abuse, including questions related to both prescription medications (opioids, tranquilizers, sedatives) and illicit drugs (e.g., heroin, cocaine, methamphetamine), drug dependency, addiction, and treatment, demographic measures of education and employment, physical health, depression, and mental health treatment. Several classification models were constructed to classify heroin use in the sample by demographics attributes and mental health characteristics (e.g., adult depression). This method addresses the following issues related to opioid dependency and addiction: (i) Identify factors related to illicit opioid use, (ii) Identify factors related to prescription opioid misuse and abuse, and (iii) Examine the relationship between prescription opioid misuse, abuse and heroin use.

## 2 METHOD

The project workflow pipeline is outlined in a readme markdown file in the project folder [22]. The steps included in the workflow were (1) Download and Extract the Data, (2) Data Cleaning and Preparation, (3) Exploratory Data Analysis, (4) Data Visualization, (5) Analysis of Classification Models for Heroin Use, and (6) Analysis of Classification Models for Prescription Opioid Pain Reliever Misuse.

### 2.1 Data

Data from the 2015 NSUH was downloaded from the Substance Abuse and Mental Health Data Archive (SAMHDA) [1] URL using the get-data.py function written to unzip the data files, extract the data as a Pandas data frame, and write the file to CSV file [4]. The dataset consists of 57,146 observations with 2,666 features representing individual-level responses from a survey of the U.S. population. According to the NSDUH codebook, sampling was weighted across states by population size for a representative distribution selected from 6,000 area segments. The sample design used five state sample size groups drawing more heavily from the eight states with the largest population (e.g., CA, FL, IL, MI, NY, OH, PA, TX) which together account for 48 percent of total U.S. population aged 12 or older. All identifying information was collapsed (e.g., age categories) and state identifiers were removed from the public use file to ensure confidentiality. The NSDUH public-use files do not include geographic location, or demographic variables related to ethnicity or immigration status. The weighted survey screening response rate was 81.94 percent and the weighted interview response rate was 71.2 percent.

### 2.2 Data Cleaning and Preparation

**2.2.1 Data Cleaning.** All steps of this analysis was completed in a python interactive notebook [16] based following examples from *Python for Data Analysis* [9]. After saving the NSDUH-2015 as a data frame object, the dataset was subset by columns to include demographic characteristics (e.g., age category, sex, marital status, education, employment status, and category of metropolitan area), measures of physical health (e.g., overall health, STDs, Hepatitis, HIV, Cancer, hospitalization), mental health (e.g., Adult Depression, Emotional Distress, Suicidal Thoughts, Plans), Suicide Attempts,

Pain Reliever Medication Use, Misuse, and Abuse (over past year, past month), Prescription Opioid Medications Taken in Past year (e.g., Hydrocodone, Oxycodone, Tramadol, Morphine, Fentanyl, Oxymorphone, Demerol, Hydromorphone), Heroin Use, Abuse (over past year, past month), Tranquilizer Use, Sedative Use, Cocaine Use, Amphetamine and Methamphetamine Use, Hallucinogen Use, Drug Treatment (e.g., Inpatient, Outpatient, Hospital, Mental Health Clinic, ER, Drug Treatment Status), and Mental Health Treatment History. A codebook was created to provide a complete list of variables included with summaries of response categories [19]. The following steps were taken to detect and remove inconsistencies in the data [13]:

- (1) Remove missing values (i.e., NaN)
- (2) Recode blanks, non-responses, or legitimate skips (e.g., 99, 991, 993) to zero
- (3) Recode dichotomous responses (e.g., Yes=1 / No=2) so that No=0
- (4) Recode categorical variables to be consistent with amount or degree (e.g., 1=low, 2=med, 3=high)
- (5) Rename selected variables for better description (e.g., Adult Major Depressive Episode Lifetime changed from AMDELT to DEPMELT)

**2.2.2 Aggregated Variables.** Because the majority of features were represented as dichotomous Yes / No variables, related features were summed to create aggregated variables. For example, overall health, STD, Hepatitis, HIV, Cancer, and hospitalization were aggregated to create a single health measure. The health measure was recoded so that higher scores indicated better health. Questions related to depression, emotional distress, and suicidal thoughts were summed to create a single variable for mental health (MENTHLTH) with scores ranging from 0 to 9. Responses to pain reliever medication use, misuse, abuse, or dependency, were aggregated to create a single variable of pain reliever misuse or abuse (PRLMISAB). All prescription painkiller medications used in the past year were summed. Similarly, all related responses were summed to create single variables for Tranquilizers, Sedatives, Cocaine, Amphetamines, Hallucinogens, Drug Treatment, and Mental Health Treatment. The target outcome of interest for classification, lifetime heroin use (i.e., “Have you ever used heroin before, at any time?”) is a dichotomous variables. The demographic characteristics and aggregated variables were subset and saved to a new data frame consisting of 2 features and 57,146 observations, which was exported to CSV file.

## 3 RESULTS

### 3.1 Exploratory Data Analysis

Of the total sample of N=57,146 respondents, 26,736 were male and 30,410 female; 6,343 individuals reported misusing pain medication at some point (570 males, 386 females), but only 956 respondents had used heroin (570 males, 386 females). Table 1 shows the raw counts of individual substance use by age group (with the sample size for each age group), listing the ten most commonly used opioid pain medications, self-reported misuse of prescription opioid pain relievers (i.e., PRL Misuse Ever), use of prescription Tranquilizers, Sedatives, and Methadone. In addition, self-reported use of illicit drugs such as heroin, cocaine, amphetamines, methamphetamine,

Hallucinogens, including LSD and Ecstasy (MDMA). This summary table shows that substance use seems to be highest for individuals between the ages of 18 to 25 and from 35 to 49 years. Of the prescription relievers, Hydrocodone use (e.g., Vicodan) was almost double the rate of Oxycodone use (e.g., Oxycodone) for each age group, and was significantly higher than any other prescription opioid medication. Use of prescription Fentanyl and Demerol, two powerful opioids, and synthetic morphines such as Oxymorphone and Hydromorphone, was very low. The rate of prescription Tranquilizer use was several orders of magnitude higher than Sedative use or Methadone use. Compared to other illicit drugs such as Cocaine, Amphetamines, Hallucinogens, heroin use was not very common in this sample. The highest rates of heroin use were seen between the ages of 18 to 49, and was lowest for respondents in the youngest age group 12 to 17, and individuals over 50.

[Table 1 about here.]

Table 2 shows the frequency of individuals reporting that they had experienced mental health issues such as depression, suicidal thoughts, whether they had received mental health treatment, received treatment from a private therapist, or believed that they needed drug treatment, but had not sought treatment, across each age category. Frequency of depression was not included for respondents between 12 to 17 years, because the survey measure was for adult depression.

[Table 2 about here.]

Figure 1 shows the proportion of individuals who reported misusing prescription opioid pain relievers and who reported using heroin. The left column of the Figure 1 shows the majority of respondents (89 percent) stated they had never misused prescription opioid pain medication or used heroin, although 10 percent reported misusing opioid pain medication at some point. The right panel of Figure 1 shows that, of those individuals who reported using heroin, the proportion who also reported misusing opioid pain medication was almost twice as large as the proportion of those who only used heroin. This is consistent with the hypothesis that misuse of prescription opioids is linked with heroin use for some individuals.

[Figure 1 about here.]

Figure 2 shows the aggregated measure of Opioid Pain Reliever misuse and abuse plotted against the aggregated measure of Heroin use (which includes misuse, abuse, lifetime use, past year use, 30 day use), with weighted regression lines grouped by size of City/Metropolitan region (from none to large). The largest proportion of the sample who report prescription opioid misuse, abuse, and heroin use is represented by observations from large metropolitan areas (red circles) with large population size. However, a small number of observations from rural or small metropolitan regions (blue and green circles) showed very high rates of prescription opioid misuse and abuse. Regression lines (i.e., line of best fit) shown are weighted by the City/Metro region attribute, with a steeper slope shown for smaller metropolitan regions than large metropolitan regions. The difference in slope may be due to the influence of the small number of outliers who had high degrees of prescription opioid misuse, and heroin use. The plot also shows a clear divide on the y-axis, which separates the sample according to high and low or no prescription

opioid misuse, although the continuum of heroin use from no, low, to high is distributed fairly evenly along the x-axis.

[Figure 2 about here.]

Figure 3 shows the pairplots of demographic features including mental health (higher scores equal to more depression), Prescription Opioid Pain Reliever (PRL) Medication (aggregated), Heroin Use (aggregated measure), and Size of City/Metropolitan region. The top row shows that the majority of the sample reported no mental health concerns, whereas a small proportion of the sample reported depression, emotional distress, or suicidal thoughts. Only few people self-described as high in depression reported low Prescription Opioid PRL misuse and abuse. The plot also reveals that prescription opioid misuse and heroin use were distributed approximately evenly for individuals reporting either low, moderate, or high levels of depression, which suggests that depression was not a factor in predicting opioid misuse. The second row shows a small number of individuals from rural areas or small cities who reported very high levels of prescription opioid misuse, although the majority of respondents misusing or abusing prescription opioid were from large metropolitan areas. As described above, the majority of respondents (about 90 percent of the sample) reported they had never misused prescription opioids. In the second row and third and fourth columns, a natural break is seen between individuals who reported high levels of prescription opioid misuse and abuse and those who reported very low or no opioid misuse. A very small proportion of the entire sample reported both misusing and abusing prescription opioids and using heroin, but this is a group of interest. The last column of the second row shows the individuals reporting high levels of opioid misuse and abuse were distributed evenly across city/metropolitan areas of different sizes, with only slightly higher numbers for small cities or rural areas. As stated above, only few participants reported using heroin, and of these, the majority were from large metropolitan areas. Finally, the sample seems to have slightly higher proportions from small and large metropolitan areas, which is likely due to weighted sampling, which drew more from heavily populated regions.

[Figure 3 about here.]

### 3.2 Classifier Models of Heroin Use

This analysis classified individuals according to whether they had ever used heroin (i.e., "Heroin Use Ever"). All classifier models were constructed using SciKit Learn [10] using an interactive python jupyter notebook [17]. The features of interest were demographic characteristics, health, mental health (adultdepression), prescription opioid misuse and abuse (PRLMISEVR, PRLMISAB, PRLANY), prescription tranquilizers use and sedatives use (TRQLZRS, SE-DATVS), use of illicit drugs (COCAINE, AMPHETMN), drug treatment (TRTMNT), and mental health treatment (MHTRTMT). The target variable was Heroin Use (HEROINEVR). Next, the dataset was split into the training set and test sets using the train-test-split() function in sklearn. Model accuracy for the training set and test set are reported, with different parameter values, and features importance.

**3.2.1 Logistic Regression Classifier.** Logistic Regression Classification is based on a linear equation that calculates the relative

weight of each feature for a categorical target or binary outcome (yes / no) [14]. The logistic regression classifier was fit to the training data in Scikit-Learn, and the model was validated on the test data. By default, the model applies L2 penalty (Ridge). The training set accuracy was 0.983 and the test set accuracy was 0.984. The parameter ‘C’ determines the strength of regularization, with higher values of C providing greater regularization. The L1 penalty (Lasso) limits the values of most coefficients to zero, creating a more interpretable model that uses only a few features. Figure 4 plots the coefficients of logistic regression classifier for heroin use with the L1 Penalty (Lasso) under different values of parameter C. The default setting, C=1.0, provides good performance for train and test sets, but the model is very likely underfitting the test data. Using a higher value of C fits a more flexible model and generally gives improved accuracy for both training and tests sets. Using a value of C=100 yielded training set accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows that the features coefficient values did not change much according to the values of parameter C, and the accuracy values were approximately the same for all values of C. Examination of the coefficients from the logistic regression classifier revealed the three features which were most closely associated with Heroin use were: Prescription Opioid Pain Reliever (PRL) Misuse ever (as predicted), Cocaine Use, and Amphetamine use, respectively.

[Figure 4 about here.]

**3.2.2 Decision Tree Classifier.** The following analysis used the *Decision Tree Classifier* package in Scikit-Learn, which only does pre-pruning. First, the decision model was build using the default setting of a fully developed tree until all leaves are pure. The random state’ features is fixed to break ties internally. Accuracy on the training set was 0.99 and test set accuracy was 0.974. Without restricting their depth, decision trees can become complex; unpruned trees are prone to overfitting and do not generalize well to new data. Limiting the depth of tree decreases overfitting, which results in lower training set accuracy, but improved performance on the test set. Next, pre-pruning was applied, with a maximum depth of 4, which means the algorithm split on four consecutive questions. Training set accuracy of the pruned tree was 0.985 and test set accuracy was 0.984. Even with a depth of 4, the tree can become a bit complex. Figure 5 shows a partial view of the decision tree classifier of heroin use (the entire tree was too wide to include as a legible Figure), and the full tree image is available in the notebook BDA-Analytics-Classifier-Heroin.ipynb [17]. The decision tree shows the top features that the algorithm split on to classify heroin use. One way to interpret a decision tree it by following the sample numbers represented at the test split for each node. The classifier algorithm selected Cocaine Use (aggregated score) as the root node of the decision tree. The branch to the left side of the tree represents samples with a score equal to or less than 1.5 (n=40956), whereas the branch to the right represents samples with a Cocaine Use score greater than 1.5 (n=1903). The second split on the right occurs for Any Prescription Opioid Pain Reliever Use (PRLANY), with n=1443 having a score less than or equal to 3.5, and n=460 respondents with a PRL score greater than 3.5. In other words, of those respondents who reported relatively high Cocaine use, a small portion also reported relatively high Prescription Opioid PRL

use. Instead of looking at the whole tree, features importance is a common summary function that rates how important each feature is for the classification decisions made in the algorithm. Each feature is assigned an importance value between 0 and 1; with a value of 1 indicating the feature perfectly predicts the target and a value of 0 meaning that the feature was not used at all. Feature importance values also always sum to 1. A feature may have a low feature importance value because another feature encodes the same information. The top two important features for classifying Heroin Use were Cocaine Use and Any Prescription Opioid PRL Use, with smaller importance given to Opioid PRL Misuse Ever and Prescription Opioid PRL Misuse and Abuse.

[Figure 5 about here.]

**3.2.3 Random Forests Classifier.** Random forests is an ensemble approach that builds many trees and averages their results to reduce overfitting. The model was build using the *Random Forest Classifier* package in Scikit-Learn. The parameters of interest for building random forests are: (a) the number of trees (n-estimators), (b) the number of data points for bootstrap sampling (n-samples), and (c) the maximum number of features considered at each node (max-features). The max-features parameter determines how random each tree is, with smaller values of max-features resulting in trees in the random forest that are very different from each other. This analysis applied a random forest consisting of 100 trees to classify Heroin Use, and the random state was set to zero. The training set accuracy was 0.999 and the test set accuracy was 0.984. Often the default settings for random forests work well, but we can apply pre-pruning as with a single tree, or adjust the maximum number of features. Feature importance for random forests is computed by aggregating the feature importance over trees in the random forest, and random forests gives non-zero importance to more features than a single tree. Typically random forests provide a more reliable measure of feature importance than the feature importance for a single tree. Figure 6 shows the feature importance of the random forests classifier for heroin use with 100 trees. Similar to the single tree, the random forest selected Cocaine Use as the most informative feature in the model, followed by Any PRL Use, which is an aggregated measure of prescription opioid medication use. Following after that, several features were tied for third place of importance, namely Education Level, Overall Health, Age Category, and Pain Reliever Misuse and Abuse. Random forests provides much of the same benefit as decision trees, while compensating for some of their shortcomings of overfitting. Single trees are still useful for visually representing the decision process.

[Figure 6 about here.]

**3.2.4 Gradient Boosting Classifier Tree.** Gradient boosting machines is another ensemble method that combines multiple decision trees for regression or classification by building trees in a serial fashion, where each tree tries to correct for mistakes of the previous one [10]. Gradient boosted regression trees use strong pre-pruning, with shallow trees of a depth of one to five. Each tree only provides a good estimate of part of the data, but combining many shallow trees (i.e., “weak learners”), the use many simple models iteratively improves performance. In addition to pre-pruning and the number of trees, an important parameter for gradient boosting is the

learning rate, which determines how strongly each tree tries to correct for mistakes of previous trees. A high learning rate produces stronger corrections, allowing for more complex models. Adding more trees to the ensemble also increases model complexity. Gradient boosting and random forests perform well on similar tasks and data; it is common to first try random forests and then include gradient boosting to attain improvements in accuracy of the learning model. This analysis used the *Gradient Boosting Classifier* from Scikit-Learn to classify Heroin Use, with the default setting of 100 trees of maximum depth of 3, and a learning rate of 0.1. The model was built on the training set and evaluated on the test set, with both training set and test set accuracy equal to 0.984. To reduce overfitting, pre-pruning could be implemented by reducing the maximum depth, or by reducing the learning rate. Figure 7 shows that the feature importance for the gradient boosting classifier tree looks similar to the feature importance for random forests, but the gradient boosting has decreased the importance of many features to zero. Again Cocaine is selected as the most informative features, followed by Any Opioid PRL Use. In addition to Prescription Opioid PRL Misuse and Abuse, the gradient boosting classifier selected Amphetamine Use as an informative feature of Heroin Use.

[Figure 7 about here.]

### 3.3 Classifier Models of Prescription Opioid Pain Reliever (PRL) Misuse

This section reports results from the same set of classification analyses described above using *Prescription Opioid Pain Reliever Misuse* (PRLMISEVR) as the target variable. Attributes related to Heroin Use were now included as features (e.g., HEROINEVR, HEROINUSE, HEROINFQY). The classifier models were built using SciKit Learn in a python notebook [18]. The dataset was split into the training set and test sets using the train-test-split function in sklearn and the target variables were designated. Model accuracy for the training set and test set are reported, for different parameter values, with feature importance.

**3.3.1 Logistic Regression Classifier.** The logistic regression classifier was fit to the training data using the L1 penalty (Lasso), using different values of the regularization parameter C, and the model was validated on the test data. Higher value of parameter C typically gives improved accuracy for both training and tests sets; however, in this case, the training set accuracy was 0.901 and test set accuracy was 0.903, and these values were consistent for all values of parameter C. Figure 8 plots the coefficients of logistic regression classifier for Prescription Opioid PRL Misuse under different values of C. As shown in Figure 8, the features with the highest coefficient values were Treatment (for substance use), Heroin Use (as predicted), as well as Cocaine and Amphetamine use. This result indicates that Prescription Opioid Misuse is positively related to Drug Treatment, meaning that respondents who reported higher levels of opioids misuse were also in treatment, but that people who were misusing opioid medications were also more likely to have used illicit drugs such as heroin, cocaine, and amphetamine.

[Figure 8 about here.]

**3.3.2 Decision Tree Classifier.** The Decision Tree Classifier package in Scikit-Learn was used to build the tree model, pre-pruning

was applied with a maximum depth of 4, which means the algorithm split on four consecutive questions. The training set accuracy of the pruned tree was 0.902 and test set accuracy was 0.902. Figure 9 shows a partial view of the decision tree classifier of prescription opioid misuse (the full tree is included in the BDA-Analytics-Classifier-PRL.ipynb notebook) [18]. As Figure 9 shows, the decision tree classifier selected Cocaine Use as the root note, that branched by the test score equal to or less than 0.5 (any Cocaine Use). At the second node, on the branch to the right n=5015 samples were further divided according to heroin use, with n=1913 having a score greater than 0.5 (any Heroin Use). At the third node on the right branch, samples were selected according to Tranquilizer medication use, with n=1419 scoring positively. On the left branch, the second node selected was Drug Treatment, with n=2844 respondents scoring positively that they had received Drug Treatment. Feature importance of the decision tree classifier selected Cocaine Use as the most informative feature for Prescription Opioid PRL Misuse. Following afterwards, Tranquilizer Use, Drug Treatment, and Heroin Use were tied for second place.

[Figure 9 about here.]

**3.3.3 Random Forests Classifier.** The Random Forest Classifier package in Scikit-Learn was used to classify Prescription Opioid PRL Misuse as the target variable, with 100 trees. The model accuracy for the training set was 0.955 and the test set accuracy was 0.896, which suggests that the model overfit the data. Figure 10 shows the feature importance of the random forests classifier for Prescription Opioid PRL Misuse. As Figure 10 shows, several features were identified as important for classifying Prescription Opioid PRL Misuse. The random forest selected Overall Health as the most informative feature in the model, followed by Cocaine Use, Education Level, Age Category, and Size of City Metropolitan region. Because of the additional features included as important, gradient boosting was performed to clarify the feature importance.

[Figure 10 about here.]

**3.3.4 Boosted Gradient Classifier.** The Gradient Boosting Classifier from Scikit-Learn was used to classify Prescription Opioid PRL Misuse, using the default setting of 100 trees, of maximum depth of 3, and a learning rate of 0.1. The model accuracy for the training set was 0.894 and accuracy for the test set was 0.893. Gradient boosting typically improves test set accuracy by using many simple models iteratively. In this case, model accuracy for gradient boosting was no better than random forests, and this is because the default parameter settings were used; further parameter tuning is needed to improve model performance. Feature importance was a primary interest for identifying features related to 'prescription opioid abuse. Figure 11 shows the feature importance for the gradient boosting classifier tree. As Figure 11 shows, several features were important for classifying prescription opioid misuse, and contrary to the random forests, gradient boosting selected Tranquilizer use as the most informative feature. Following closely in importance were Heroin Use and Age Category. Tied for fourth place were Cocaine Use and Treatment, with Mental Health (depression) coming in fourth in terms of feature importance. This result illustrates that several features are important for understanding Prescription Opioid Misuse, and the relations among features may be complex.

[Figure 11 about here.]

## 4 DISCUSSION

The results show that rates of prescription opioid use, misuse, and abuse are much higher than use of illicit opioids such as heroin and fentanyl. The use of Hydrocodone (Vicodan) was double the rate of Oxycodone use (Oxycodone) across almost all age groups. The use of traditional prescription opioids was greater than reported use of synthetic opioids. Illicit drug use was highest for respondents between the ages of 18 to 25. In terms of mental health, more individuals between 18 to 25 years reported experiencing a major depressive episode (in adulthood) than any other age group. In terms of the so-called *treatment gap*, almost twice as many respondents between 18 to 25 years who felt a need for substance use treatment, had not received treatment, than younger individuals between 12 to 17 years. The large majority of respondents (approximately 90 percent) had not misused prescription opioid pain relievers or used heroin. However, of those individuals who reported misusing prescription opioid pain relievers, almost twice as many had also used heroin than had not (see Figure 1), which partially supports the hypothesis that prescription opioid use is associated with use of illicit opioids such as heroin. Prescription opioid misuse and heroin use was also higher in large metropolitan areas than smaller cities or rural areas, but a small portion of individuals in non-metropolitan regions reported very high levels of prescription opioid misuse. These data points may represent outliers, but a large sample would allow for analysis of how opioid misuse and addiction differ for smaller rural regions versus large urban areas.

### 4.1 Comparison of Classifier Models

Several classifier algorithms were used to identify relevant features for predicting heroin use and prescription opioid misuse. Comparing the performance of different algorithms is helpful for selecting the best model. Test set accuracy was comparable across models for both Heroin Use (0.98) and Prescription Opioid PRL Misuse (0.89-0.90). Logistic Regression provided the feature coefficients for different values of the regularization parameter C. The Decision Tree classifier provided an easy to use, interpretable visual of the decisions involved at each step of classification. Random forests provides a more reliable indication of features importance than a single tree, whereas the gradient boosting classifier included additional tuning parameter for a more powerful model and more interpretable analysis of feature importance. Each classifier method provides a different level of analysis. For classifying heroin use, the logistic regression classifier showed that Prescription Opioid PRL Misuse had the highest coefficient value, but the tree-based classifiers each identified Cocaine Use as the most informative feature for predicting heroin use. For classifying Prescription Opioid PRL Misuse, logistic regression showed that Treatment had the highest coefficient value, but the tree based models each differed in selecting the most important features. Decision trees indicated that Cocaine Use was most informative, the random forests classifier selected health as the most important feature, and the gradient boosting model selected Tranquillizer use as most informative of prescription opioid PRL misuse. The different model each have

their advantages and limitations, logistic regression provides the coefficients, but random forests and gradient boosting are helpful for identified sets of important features.

### 4.2 Study Limitations

The main goal of this project was to identify features relevant for predicting opioid addiction by classifying cases according to heroin use. Only a small proportion of the sample reported having used heroin, and scores for mental health issues were very low. A limitation of survey data is that responses may be biased by under-reporting or minimizing the use of illicit or illegal substances. People may also be reluctant to disclose mental health issues or health problems (e.g., STDs, HIV status, suicide attempts). It is possible that this sample is representative of the frequency of opioid use and misuse in the larger population. Recent statistics from the CDC show that heroin use has increased among most demographics groups, with an average estimated rate of approximately 2.6 percent between 2011-2013 [7]. The rate of heroin use reported in the NSDUH-2015 sample was 1.6 percent. Therefore, it seems that the actual rate of heroin use in the U.S. population may not be accurately reflected in this sample. Another limitation is that the project dataset was constructed as a subset of features from the NSDUH-2015 data. Ninety attributes out of 2666 features in the original data were selected, and many features were combined to create aggregated variables for health, mental health, prescription opioid misuse and abuse, drug treatment, mental health treatment. Future research could include a more comprehensive selection of features to identify the set of features relevant for predicting opioid dependency and addiction. An important challenge for making sense of big data is developing analytic tools adequate to handle large volumes of data.

### 4.3 Extension to Big Data

A general tenet of big data is that, “More data is always better.” The methods used in this project could be extended to better approximate big data for predicting opioid use in the following ways: (1) Include a larger selection of features from the attributes in the NSDUH-2015 dataset; (2) Include survey data from previous years (e.g., 2005-2015) for a larger sample; and (3) Obtain a broader sample from the population of patients who are taking prescribed opioid medications. The most immediate step would be to include additional features for use with the classifier models. Additional data from the NSDUH was downloaded from previous years (2012 to 2014); preliminary examination of the data revealed inconsistencies in questions and prescription opioid medications that would need to be resolved in order to combine data from multiple years. Data cleaning can be a time consuming process, but important for obtaining usable data. Unfortunately, owing to constraints of time for completing the project, it was not possible to integrate data from previous years into the project dataset. In working with big data, there are several steps involved in the consolidation of data from multiple sources into a single dataset (in addition to data cleaning), which include extraction, integration, and aggregation of features [13]. A future study could integrate data from different years, using a broader set of features, with more inclusive sample

representative of the larger population, and integrate data from multiple sources.

#### 4.4 Opioid Addiction and Epidemic Spreading

Drug addiction has many similar characteristics to other chronic medical illnesses, but there are unique challenges to the treatment of addiction [8, 23]. In drug rehabilitation treatment programs, patients undergo intense detoxification that reduces their drug tolerance, but are then released back into the environments associated with their drug use, putting them at high risk for relapse and potential drug overdose [6]. If the prescription opioid crisis is a genuine epidemic, we must consider the process of spreading or diffusion of contagion. Epidemic spreading is a dynamic process based on networks of direct person-to-person contact and indirect exposure via transportation pathways [2]. Epidemics are quantified in terms of the proportion of the population infected, those yet to be infected, and the rate of transmission. Potentially everyone is at risk of becoming dependent or addicted to prescription medications or illicit opioids. In terms of the opioid epidemic, rather than labeling persons as infected or uninfected, it is more useful to consider people as either susceptible to dependence and addiction or less susceptible. Furthermore, the structure of the contact network can influence epidemic spreading [12]. For example, in the case of simple contagion, weak ties among acquaintances or infrequent associations provide shortcuts between distant nodes that reduce distance within the network [?] which can facilitate the spread of contagion, or in this case drug use. Furthermore, contact networks for drug use may have “small world” properties where a small number of nodes have a high number of connections that can rapidly transmit contagion throughout the network [?]. Network analysis may help to identify the underlying structure of the contact network of opioid use, to examine pathways and points of contact in the misuse and abuse of prescription opioid medications. According to a classical conditioning model of addiction, situational cues or events can elicit a motivational state underlying relapse to drug use. Addictive behavior can be also be reinstated after extinction of dependency by exposure to drug-related cues or stressors in the environment [15]. Future research could use social network modeling to explore how drug dependency and addiction are subserved by patterns of social interaction.

### 5 CONCLUSION

This project compared several classification algorithms to predict heroin use and prescription opioid misuse and abuse. The results provided partial support for the hypothesis that prescription opioid misuse is associated with the use of illicit opioids such as heroin. Several features were identified as important for classifying heroin use, including Cocaine Use, Amphetamine Use, and any prescription opioid medication use. In regards to predicting heroin use, it appears the use of other illicit drugs such as Cocaine and Amphetamine was perhaps more informative than any prescription opioid use or misuse. Heroin use was selected as important for classifying prescription opioid pain reliever misuse, but additional factors also played a role, including tranquilizer use, age category, overall health, cocaine use. Substance treatment had the largest regression coefficient, suggesting that people who are misusing

prescription opioid pain medication are also more likely to be in drug treatment programs. The direction of these effects cannot be determined owing to the nature of the analyses. On the one hand individual misusing or abusing prescription opioids may also be using heroin. Alternatively, individuals with a susceptibility for opioid use may be equally likely to have used heroin and also to have misused prescription opioids. A general conclusion is that of those individuals who reported misusing prescription opioid medications, twice as said they had used heroin than reported they had not used heroin. The results do not provide sufficient evidence to rule out alternative hypotheses. Given the relatively low rates of opioid and heroin in this sample, additional evidence is needed to resolve this question. The study can provide information to raise awareness about the risk factors for prescription opioid addiction and may help reduce opioid overdose deaths.

### ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, the Teaching Assistants, Juliette Zurick, Miao Jiang, Hungri Lee, Grace Li, Saber Sheybani Moghadam, and others who helped to improve this project and report.

### REFERENCES

- [1] Substance Abuse, Center for Behavioral Health Statistics Mental Health Services Administration, and Quality. 2016. *National Survey on Drug Use and Health (NSDUH) 2015*. Online data archive. United States Department of Health and Human Services, Ann Arbor, MI. <https://doi.org/10.3886/ICPSR50011.v1>
- [2] Vittoria Colizza, Alain Barrat, Marc Barthélémy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7 (2006), 2015–2020. <https://doi.org/10.1073/pnas.0510525103> arXiv:<http://www.pnas.org/content/103/7/2015.full.pdf>
- [3] Centers for Disease Control and Prevention. 2017. Prescription Opioid Overdose Data. online. (Oct. 2017). <https://www.cdc.gov/drugoverdose/data/overdose.html>
- [4] hd1 and yoavram. 2016. Python: Download Returned Zip file from URL. Online. (Feb. 2016). <https://stackoverflow.com/questions/9419162/python-download-returned-zip-file-from-url> Stackoverflow.com.
- [5] M. Herland, T. M. Khoshgoftaar, and R. Wald. 2014. A review of data mining using big data in health informatics. *Journal Of Big Data* 1, 2 (2014). <https://doi.org/10.1186/2196-1115-1-2>
- [6] K. Johnson, A. Isham, D.V. Shah, and D.H. Gustafson. 2011. Potential Roles for New Communication Technologies in Treatment of Addiction. *Current psychiatry reports*, (2011). <https://doi.org/10.1007/s11920-011-0218-y>
- [7] Rose A Judd, Noah Aleshire, Jon E. Zibbell, and R. Matthew Gladden. 2016. *Increases in Drug and Opioid Overdose Deaths, United States, 2000–2014*. techreport 64(50). Centers for Disease Control and Prevention, Atlanta, GA. <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm> Morbidity and Mortality Weekly Report (MMWR).
- [8] Lisa A. Marsch. 2012. Leveraging technology to enhance addiction treatment and recovery. *Journal of Addictive Diseases* 31, 3 (2012), 313–318. <https://doi.org/10.1080/10550887.2012.694606>
- [9] Wes McKinney. 2017. *Python for Data Analysis*. O'Reilly Media Inc., Sebastopol, CA. <https://github.com/wesm/pydata-book>
- [10] Andreas C. Müller and Sarah Guido. 2017. *Introduction to Machine Learning*. O'Reilly, Sebastopol, CA. [https://github.com/amueller/introduction\\_to\\_ml\\_with\\_python/](https://github.com/amueller/introduction_to_ml_with_python/)
- [11] National Institute on Drug Abuse (NIDA). 2017. *Overdose Death Rates*. Summary. National Institutes of Health (NIH), Washington D.C. <https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates>
- [12] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* 86 (Apr 2001), 3200–3203. Issue 14. <https://doi.org/10.1103/PhysRevLett.86.3200>
- [13] E. Rahm and H. Hai Do. 2000. *Data cleaning: Problems and current approaches*. techreport 23(4). Bulletin of the Technical Committee on Data Engineering, 1730 Massachusetts Avenue, Washington D.C. <https://s3.amazonaws.com/academia.edu.documents/41858217/A00DEC-CD.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511155930&Signature=VWRM7u4KWTp6ZxX5jB%2Bh6wMCbpg%3D&response-content-encoding=base64&response-content-type=application/pdf>

- disposition=inline%3B%20filename%3DAutomatically\_extracting\_structure\_.from.pdf#page=5
- [14] Sebastian Raschka and Vahid Mirjalili. 2017. *Python Machine Learning, Second Edition*. Packt, Birmingham, UK. <https://github.com/rasbt/python-machine-learning-book-2nd-edition>
  - [15] Yavin Shaham, Uri Shalev, Lin Lu, Harriet de Wit, and Jane Stewart. 2003. The reinstatement model of drug relapse: history, methodology and major findings. *Psychopharmacology* 168, 1 (01 Jul 2003), 3–20. <https://doi.org/10.1007/s00213-002-1224-x>
  - [16] S.M. Shiverick. 2017. BDA Project Data. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project.ipynb>
  - [17] S.M. Shiverick. 2017. Classification Models of Heroin Use. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-Heroin.ipynb> Interactive Python Jupyter Notebook.
  - [18] S.M. Shiverick. 2017. Classification Models of Prescription Opioid Pain Relievers Misuse. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-PRL.ipynb> Interactive Python Jupyter Notebook.
  - [19] S.M. Shiverick. 2017. Project Codebook for Data Variables from NSDUH-2015. github. (Nov. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/project-data-codebook.txt>
  - [20] S. M. Shiverick. 2017. Exploratory Data Analysis. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb>
  - [21] S. M. Shiverick. 2017. Project Data Visualization. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb>
  - [22] S. M. Shiverick. 2017. Project Workflow Pipeline. Github. (Dec. 2017). <https://github.com/bigdata-i523/hid335/blob/master/project/readme.md>
  - [23] J. Swendsen. 2016. Contributions of mobile technologies to addiction research. *Dialogues Clinical Neuroscience* 18, 2 (June 2016), 213–221. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969708/>
  - [24] Jake VanderPlas. 2017. *Python Data Science Handbook*. O'Reilly Media Inc., Sebastopol, CA. <https://jakevdp.github.io/PythonDataScienceHandbook/>
  - [25] Upkar Varshney. 2013. Smart medication management system and multiple interventions for medication adherence. *Decision Support Systems* 55, 5 (May 2013), 538–551. <https://doi.org/10.1016/j.dss.2012.10.011>
  - [26] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. 2014. Medication-Assisted Therapies: Tackling the Opioid-Overdose Epidemic. *New England Journal of Medicine* 370, 22 (2014), 2063–2066. <https://doi.org/10.1056/NEJMmp1402780> arXiv:<http://dx.doi.org/10.1056/NEJMmp1402780> PMID: 24758595.

## A CODE REFERENCES

All code, notebooks, files, and folders for this project can be found in the i523/hid335/project github repository: <https://github.com/bigdata-i523/hid335/tree/master/project>. An outline of the workflow pipelines was included as a readme.md markdown file [22].

### A.1 Download and Extract Data

The get-data.py function was written to download the data, unzip the data files, extract the data, and write the NSDUH-2015 dataset to CSV file [4].

### A.2 Data Cleaning and Preparation

Data cleaning and preparation steps was conducted using an interactive python Jupyter Notebook [16] based on examples in Python for Data Analysis [9] and the Python Data Science Handbook [24].

### A.3 Exploratory Data Analysis

Exploratory Data Analysis of the NSDUH-2015 dataset was conducted using an interactive python notebook [20] based on examples from Python for Data Analysis [9], and the Python Data Science Handbook [24].

## A.4 Data Visualization

Several plots and graphs were constructed in a Data Visualization interactive python notebook [21] using Matplotlib and Seaborn python visualization packages [9, 24].

## A.5 Classification Algorithms

Machine learning classification models were constructed using SciKit Learn [10, 14] in two separate Jupyter Notebooks, one for classifier models of Heroin Use as the target variable [17], and another for classifier models of Prescription Opioid PRL Misuse as the target [18].

## B ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

### B.1 Assignment Submission Issues

DONE:

Do not make changes to your paper during grading, when your repository should be frozen.

### B.2 Uncaught Bibliography Errors

DONE:

Missing bibliography file generated by JabRef

DONE:

Bibtex labels cannot have any spaces, \_ or & in it

DONE:

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

### B.3 Formatting

DONE:

Incorrect number of keywords or HID and i523 not included in the keywords

DONE:

Other formatting issues

### B.4 Writing Errors

DONE:

Errors in title, e.g. capitalization

DONE:

Spelling errors

DONE:

Are you using a and the properly?

DONE:

Do not use phrases such as *shown in the Figure below*. Instead, use *as shown in Figure 3*, when referring to the 3rd figure

DONE:

Do not use the word *I* instead use *we* even if you are the sole author

DONE:

Do not use the phrase *In this paper/report we show* instead use *We show*. It is not important if this is a paper or a report and does not need to be mentioned

DONE:

If you want to say *and* do not use & but use the word *and*

DONE:

Use a space after . , :

DONE:

When using a section command, the section title is not written in all-caps as format does this for you

\section{Introduction} and NOT \section{INTRODUCTION}

## B.5 Citation Issues and Plagiarism

DONE:

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

DONE:

Claims made without citations provided

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material

## B.6 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use “quotes”, instead of ” ”

DONE:

To emphasize a word, use *emphasize* and not “quote”

DONE:

When using the characters & # % \_ put a backslash before them so that they show up correctly

DONE:

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

DONE:

If you see a ffigure and not a figure in text you copied from a text that has the fi combined as a single character

## B.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

DONE:

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

DONE:

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

DONE:

Do not artificially inflate your paper if you are below the page limit

## B.8 Details about the Figures and Tables

DONE:

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

DONE:

Do use *label* and *ref* to automatically create figure numbers

DONE:

Wrong placement of figure caption. They should be on the bottom of the figure

DONE:

Wrong placement of table caption. They should be on the top of the table

DONE:

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, jpg

DONE:

Do not submit eps images. Instead, convert them to PDF

DONE:

The image files must be in a single directory named "images"

DONE:

In case there is a powerpoint in the submission, the image must be exported as PDF

DONE:

Make the figures large enough so we can read the details. If needed make the figure over two columns

DONE:

Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to float. For this class, you should place all figures at the end of the report.

DONE:

In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

DONE:

Remove any figure that is not referred to explicitly in the text (As shown in Figure ..)

DONE:

Do not use `textwidth` as a parameter for `includegraphics`

DONE:

Figures should be reasonably sized and often you just need to add `columnwidth`

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}  
re
```

## LIST OF FIGURES

1	Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin	13
2	Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size	14
3	Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area	15
4	Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C)	16
5	Decision Tree Classification of Heroin Use (Partial View)	17
6	Feature Importance for Random Forests Classifier for Heroin Use	18
7	Feature Importance for Gradient Boosting Classifier for Heroin Use	19
8	Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty	20
9	Decision Tree for Prescription Opioid (PRL) Misuse	21
10	Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse	22
11	Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse	23

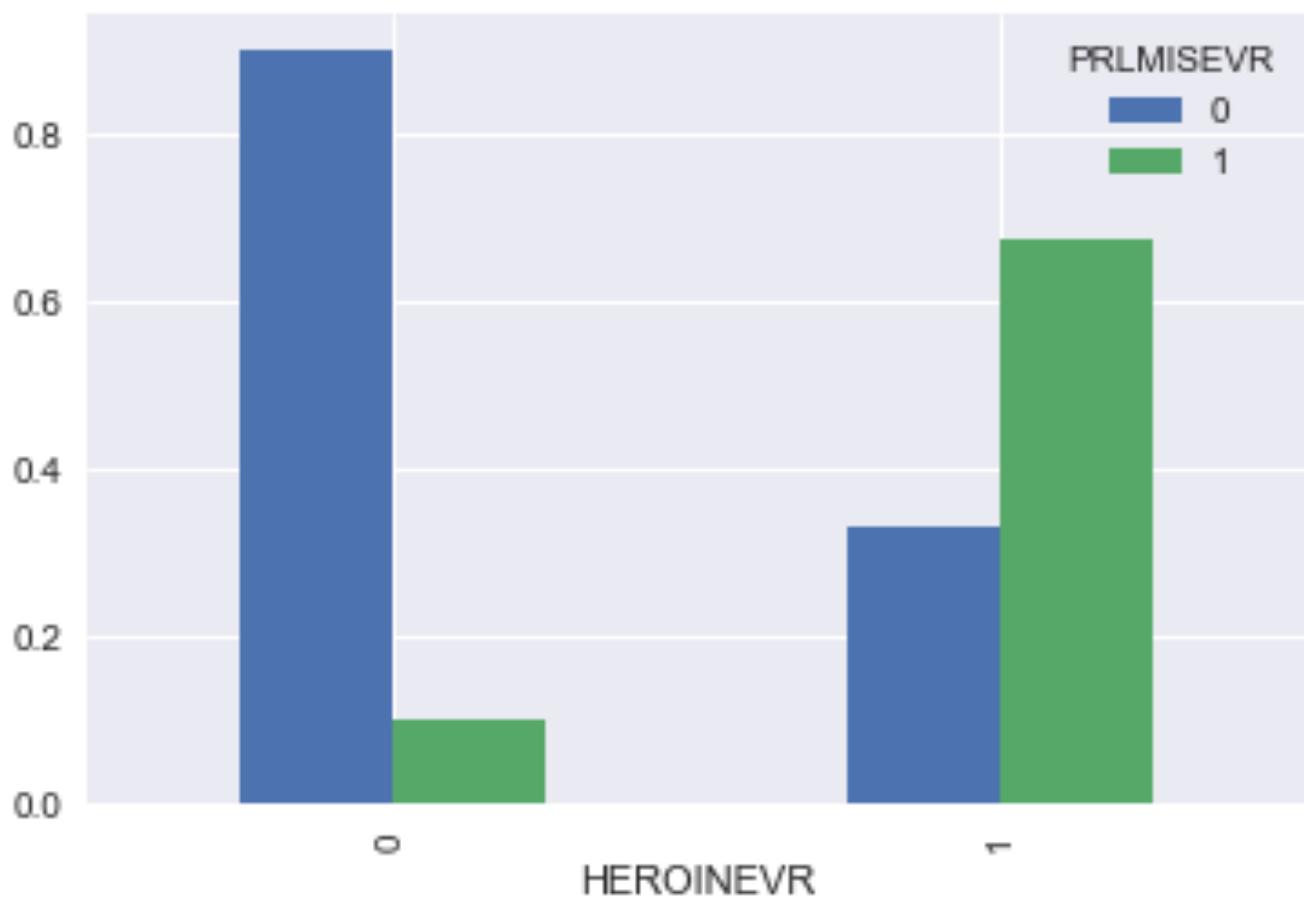


Figure 1: Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin

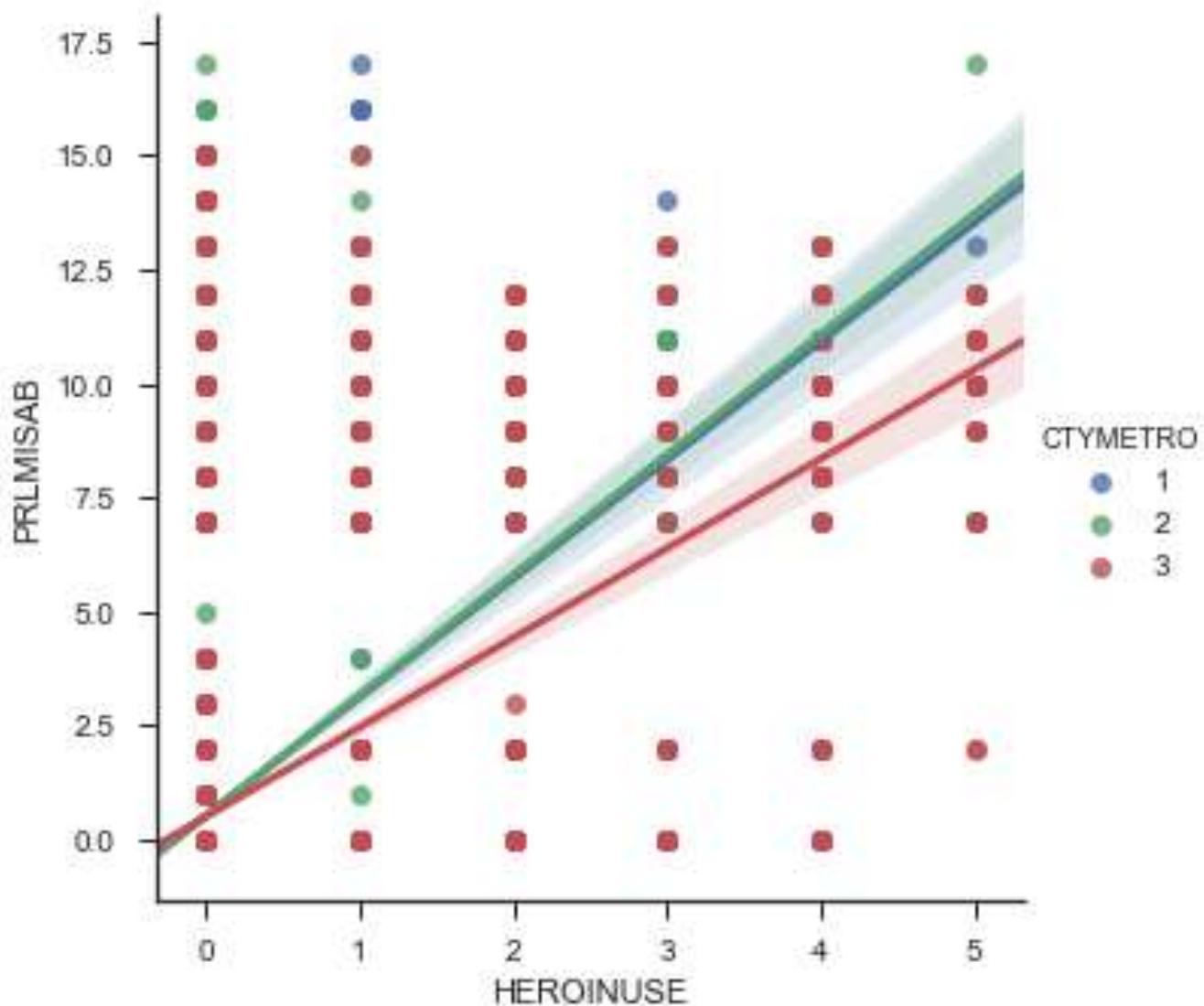


Figure 2: Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size

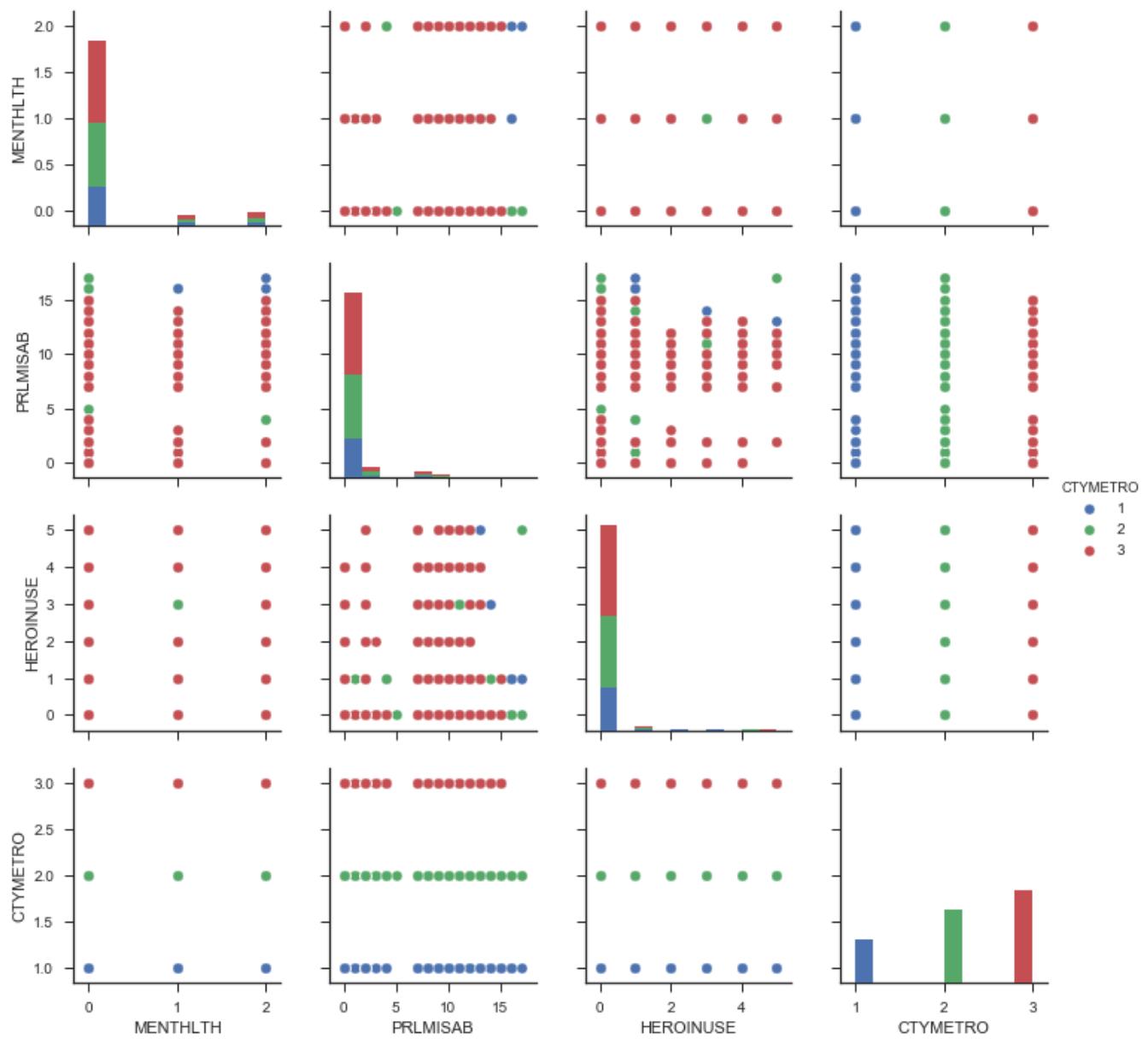


Figure 3: Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area

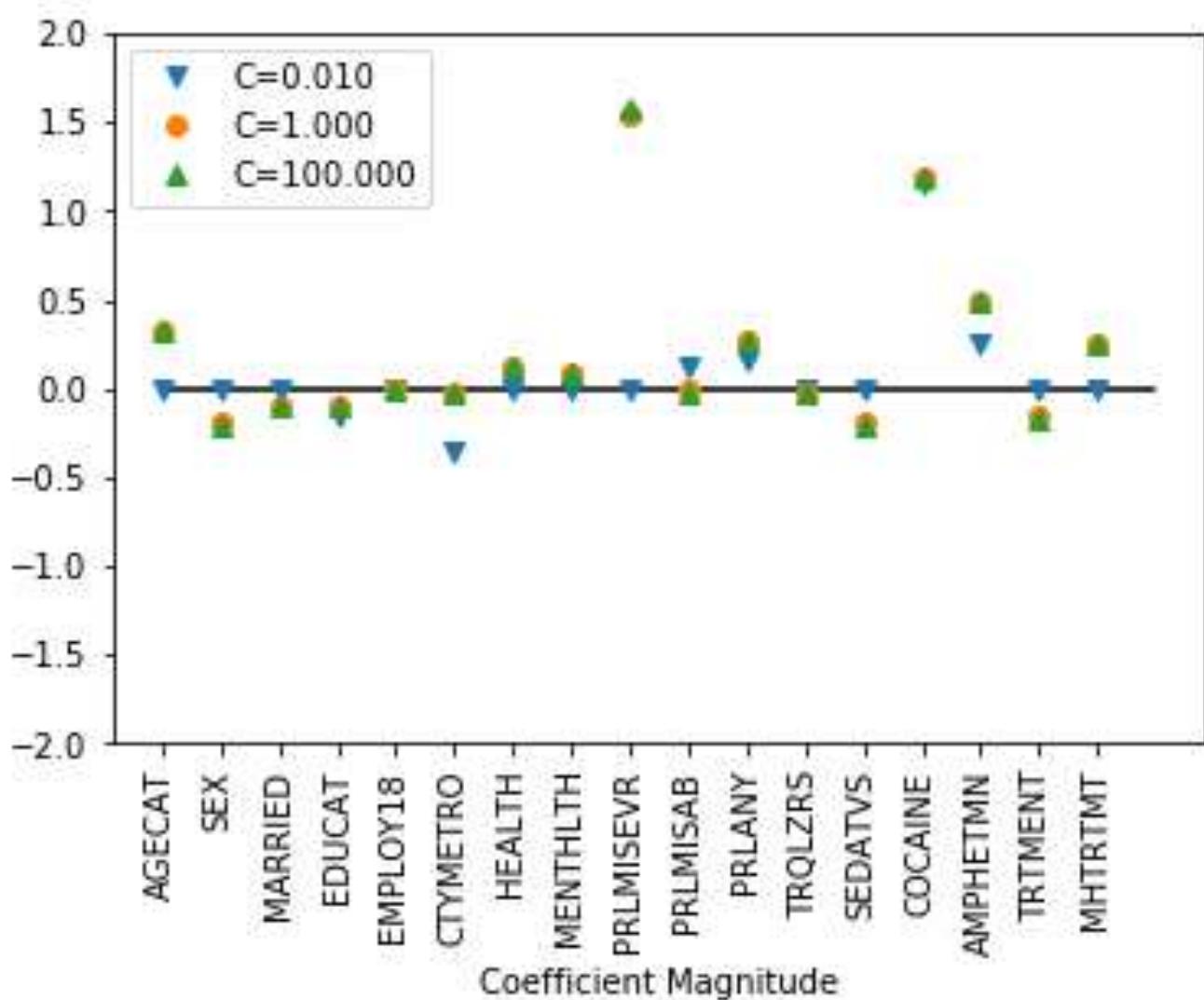


Figure 4: Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C)

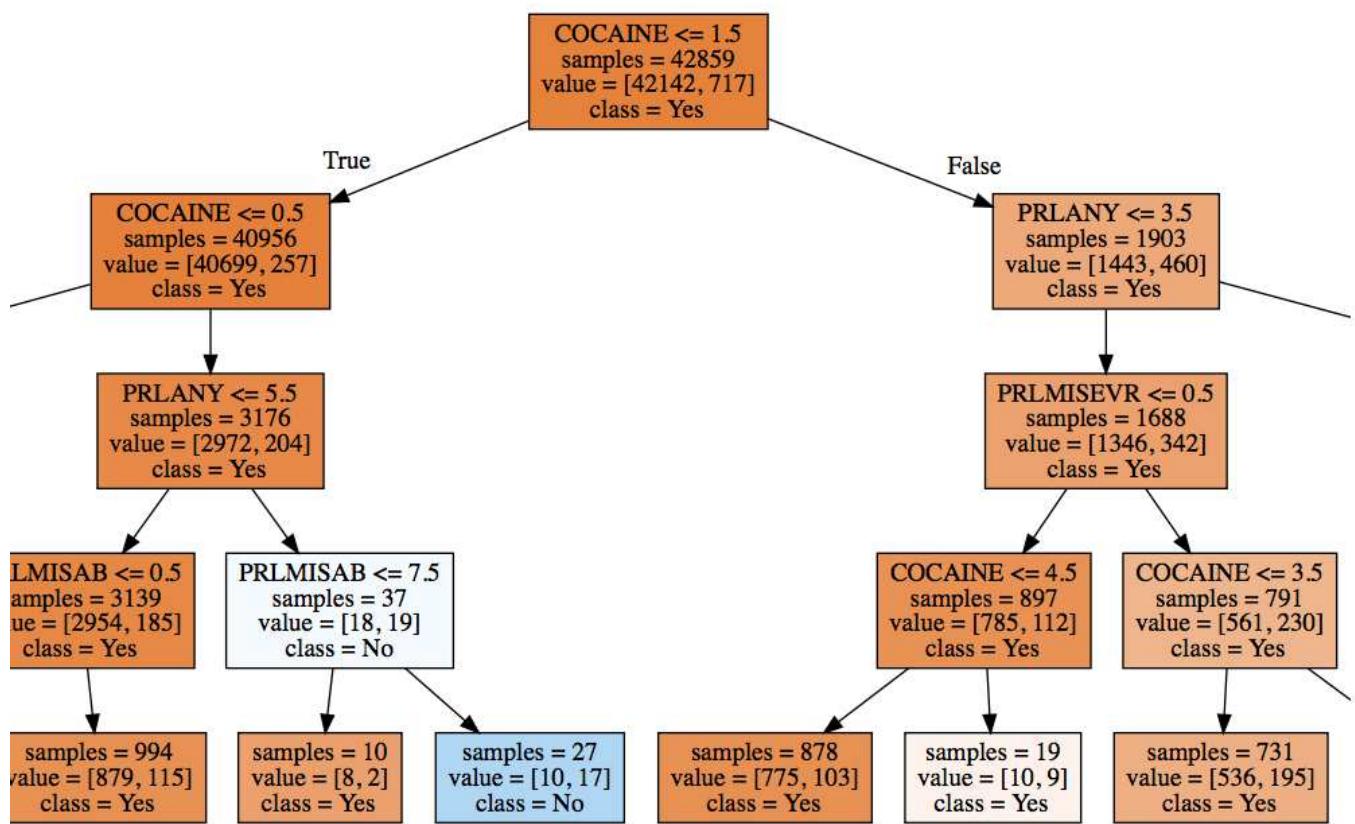


Figure 5: Decision Tree Classification of Heroin Use (Partial View)

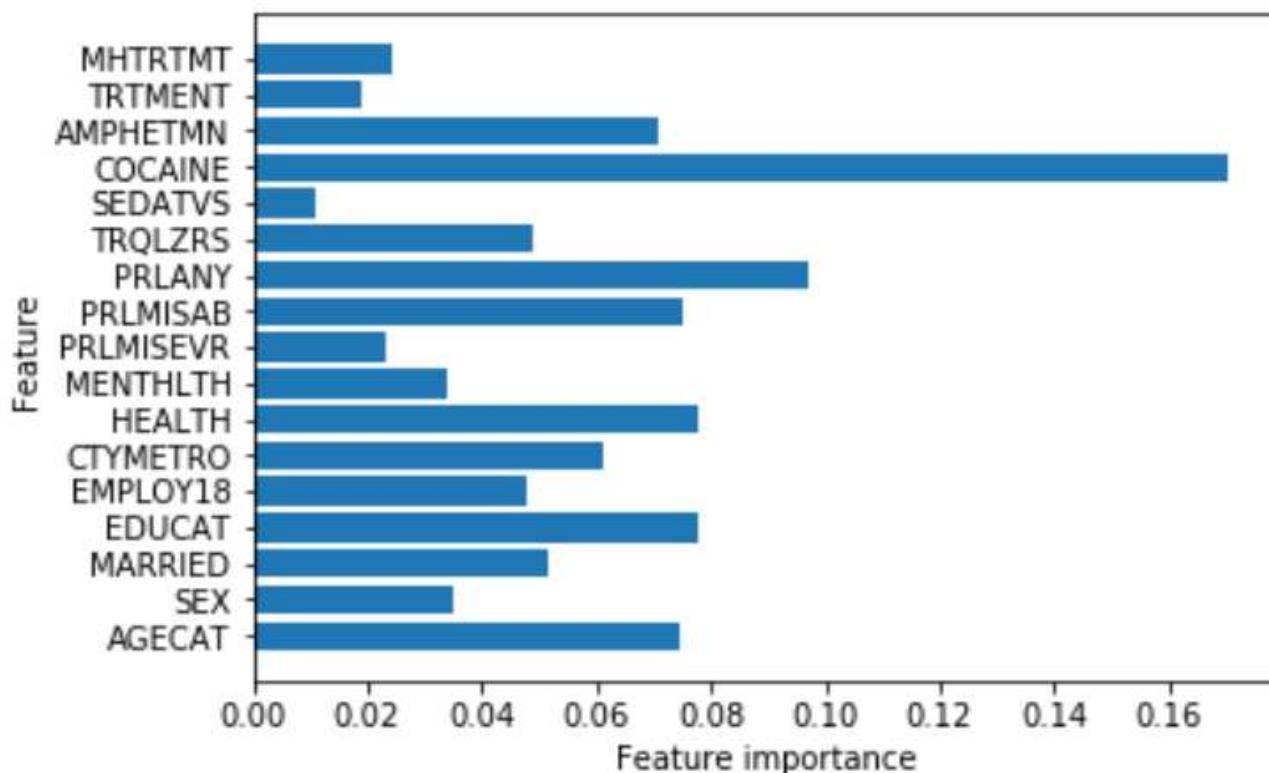


Figure 6: Feature Importance for Random Forests Classifier for Heroin Use

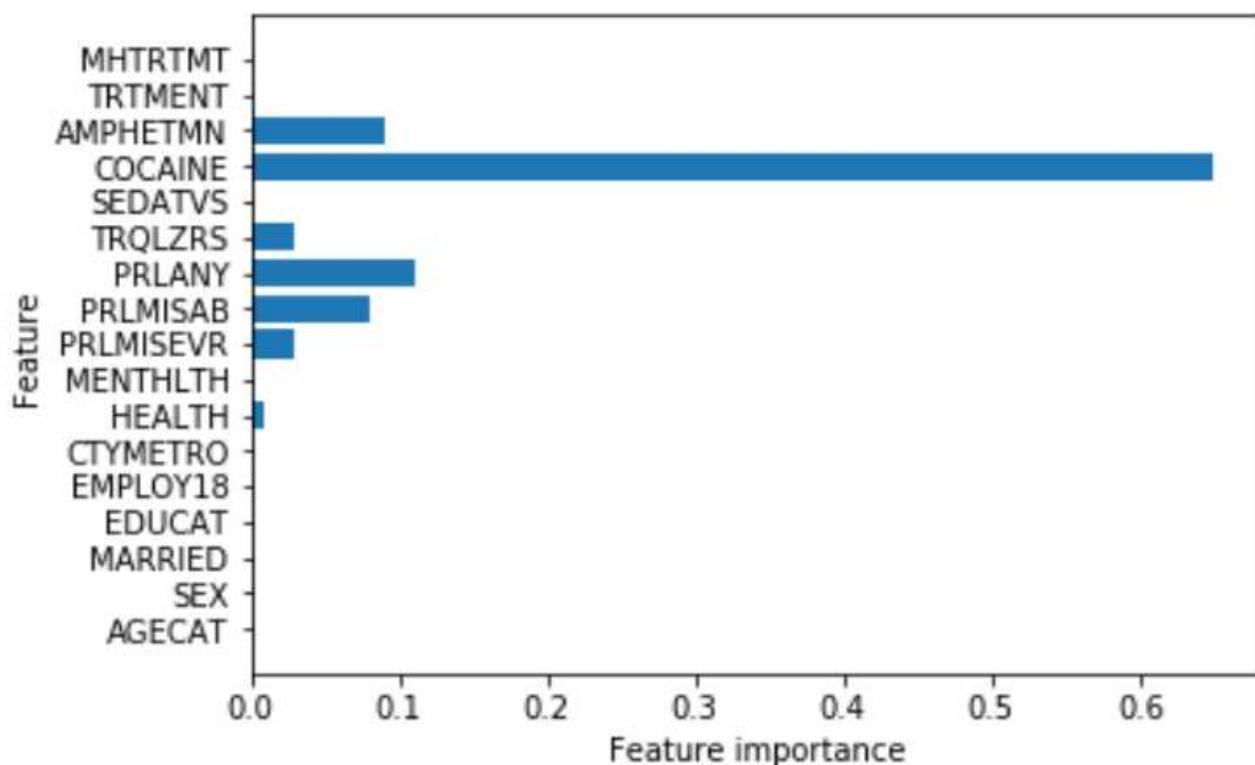


Figure 7: Feature Importance for Gradient Boosting Classifier for Heroin Use

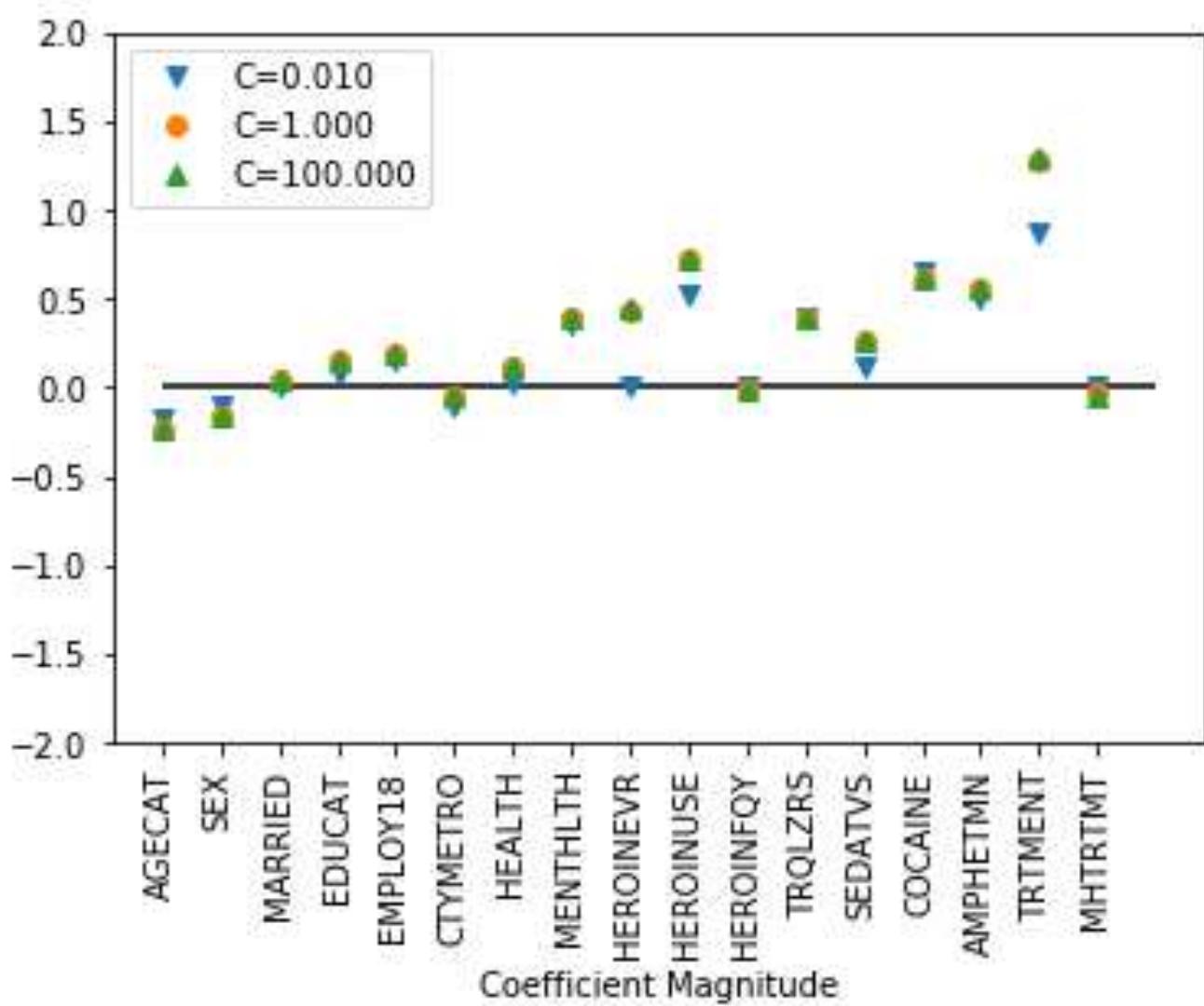


Figure 8: Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty

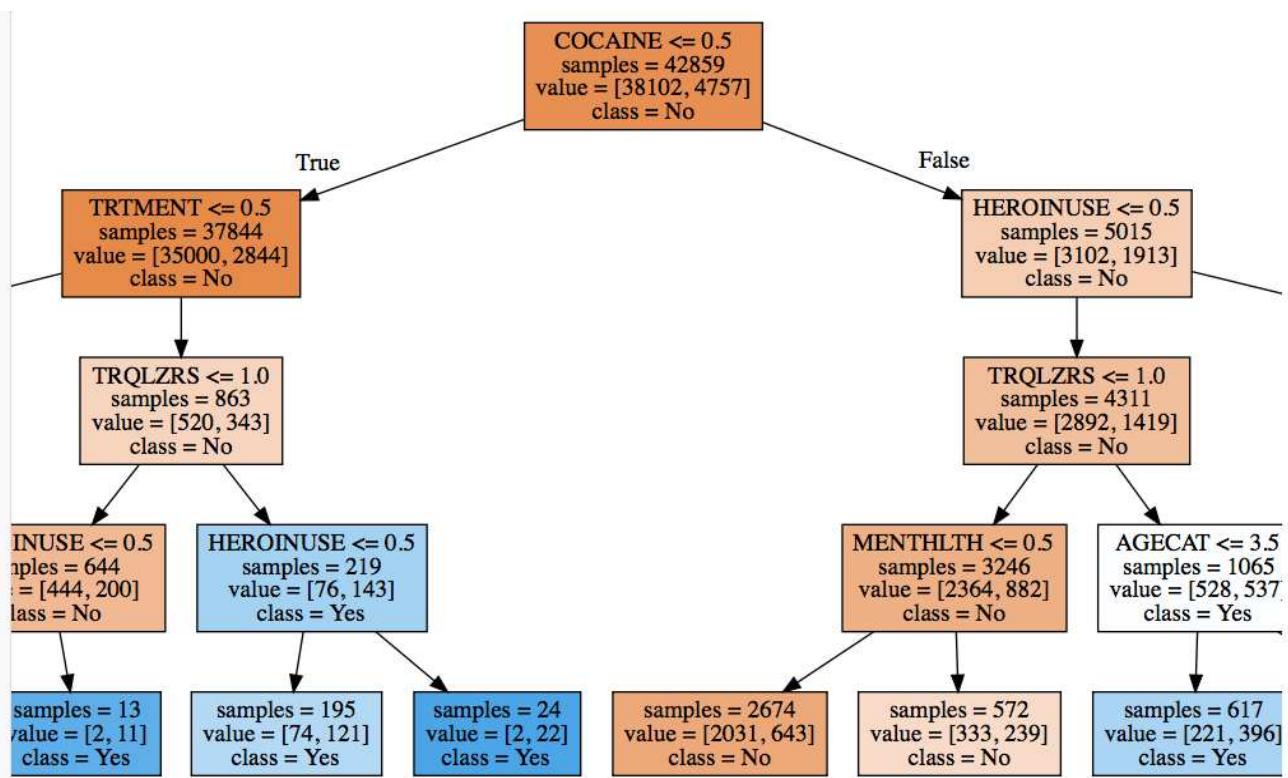


Figure 9: Decision Tree for Prescription Opioid (PRL) Misuse

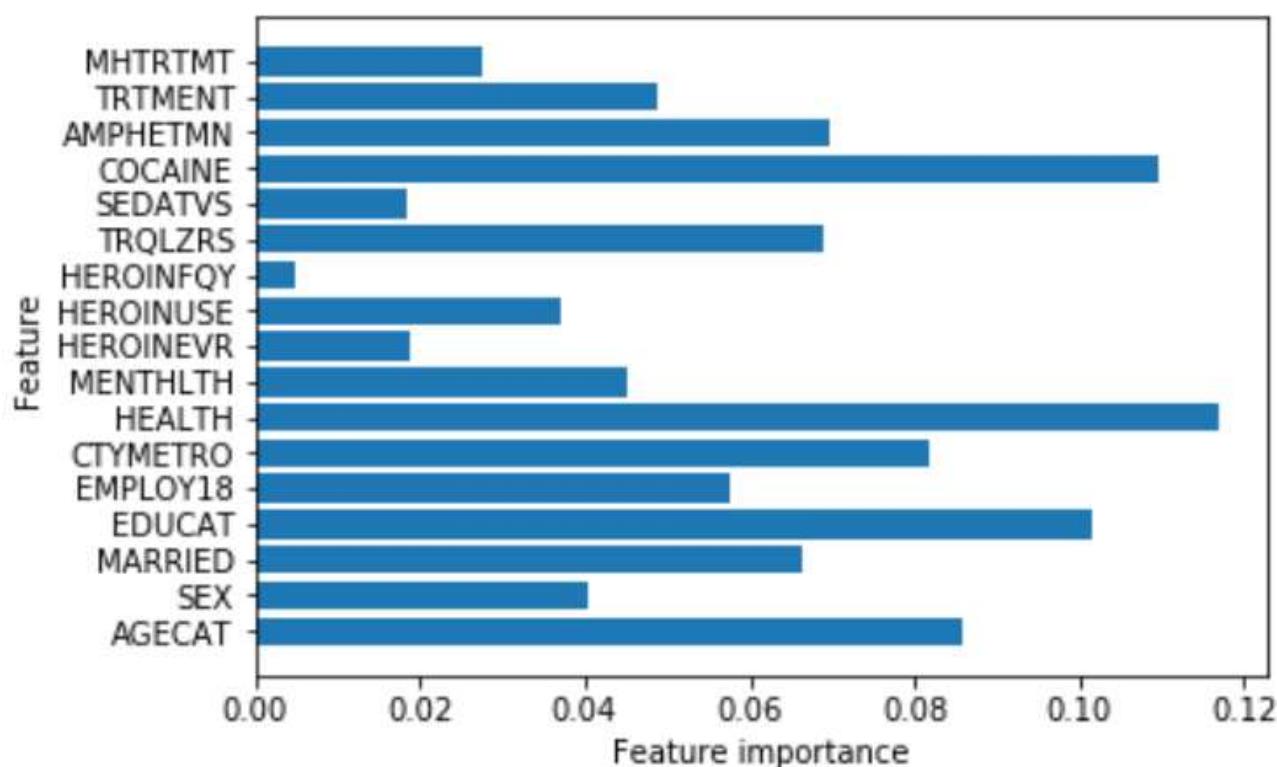


Figure 10: Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse

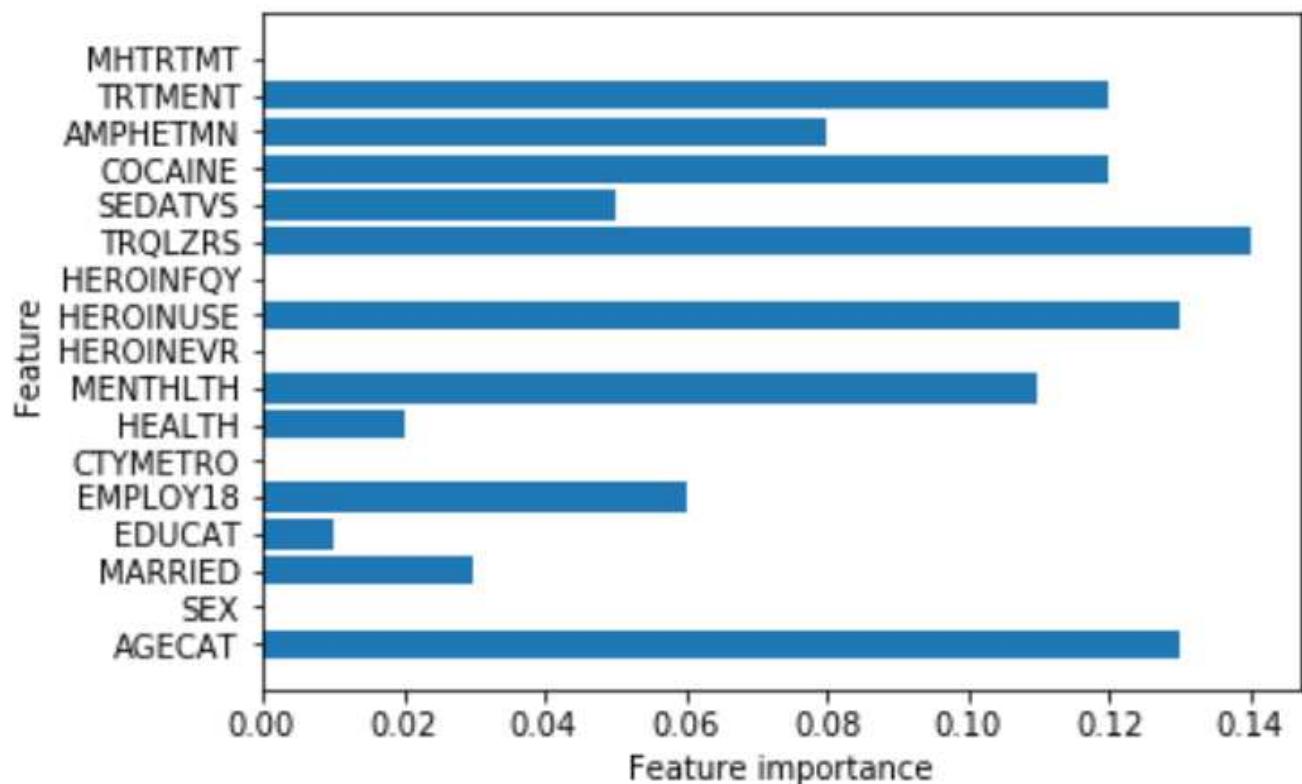


Figure 11: Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse

LIST OF TABLES

1	Substance Use by Age Group Counts - NSDUH 2015 [1]	25
2	Frequency Table of Mental Health Issues and Treatment NSDUH 2015 [1]	25

**Table 1: Substance Use by Age Group Counts - NSDUH 2015 [1]**

Age Group	12-17	18-25	26-34	35-49	50+
Sample Size	13585	14553	9084	11169	8755
Oxycodone	545	1632	1132	1345	1044
Hydrocodone	831	2936	2233	2781	2103
Tramadol	241	753	654	829	734
Morphine	251	431	236	313	286
Fentanyl	28	97	81	96	86
Demerol	26	74	49	64	71
Buprenorphine	43	197	167	124	51
Oxymorphone	46	88	57	47	41
Hydromorphone	24	94	107	118	81
PRL Misuse Ever*	798	2127	1475	1343	600
Tranquilizers	405	1469	1064	1405	1153
Sedatives	204	242	157	256	226
Methadone Ever	32	83	96	71	46
Heroin Use Ever*	22	261	259	250	164
Cocaine Use Ever	109	1645	1626	1954	1406
Amphetamines Ever	932	1836	627	383	164
Methamphetamine	42	481	700	898	492
Hallucinogens	450	2660	2020	2127	1197
LSD Use Ever	190	1114	874	1442	907
Ecstasy (MDMA)	199	1867	1403	947	149

**Table 2: Frequency Table of Mental Health Issues and Treatment NSDUH 2015 [1]**

Age Group	12-17	18-25	26-34	35-49	50+
In Hospital Overnight	730	1149	821	890	1173
Adult Depression	0	2413	1395	1766	967
Mental Health Treatment					
Private Therapist	0	592	434	554	311
Treatment Gap*	469	931	321	239	90

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "granovetter73"
Warning--I didn't find a database entry for "watts98"
Warning--page numbers missing in both pages and numpages fields in herland14
Warning--no number and no volume in johnson11
Warning--page numbers missing in both pages and numpages fields in johnson11
(There were 5 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.38.39] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Label `tab:freq' multiply defined.
p.8 L932 : [granovetter73] undefined
p.8 L936 : [watts98] undefined
Missing character: ""
There were undefined citations.
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
```

The anchor of a bookmark and its parent's must not be the same. Added a new anchor.  
There were multiply-defined labels.  
Typesetting of "report.tex" completed in 1.5s.

```
=====
Compliance Report
=====
```

```
name: Sean Shiverick
hid: 335
paper1: 10/25/17 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
25
wc 335 project 25 8447 report.tex
wc 335 project 25 9533 report.pdf
wc 335 project 25 1078 report.bib
```

```
find "
```

---

```
passed: True
```

---

```
find footnote
```

---

```
passed: True
```

---

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

---

```
find input{format/final}
```

---

```
passed: False

floats
-----
355: \begin{table}
358: \label{tab:freq}
403: \begin{table}
406: \label{tab:freq}
435: \begin{figure} [!ht]
436: \centering\includegraphics[width=\columnwidth]{images/Figure1.pdf}
}
439: \label{f:Figure1}
463: \begin{figure} [!ht]
464: \centering\includegraphics[width=\columnwidth]{images/Figure2.pdf}
}
467: \label{f:Figure2}
502: \begin{figure} [!ht]
503: \centering\includegraphics[width=\columnwidth]{images/Figure3.pdf}
}
506: \label{f:Figure3}
552: \begin{figure} [!ht]
553: \centering\includegraphics[width=\columnwidth]{images/Figure4.pdf}
}
556: \label{f:Figure4}
601: \begin{figure} [!ht]
602: \centering\includegraphics[width=\columnwidth]{images/Figure5.pdf}
}
604: \label{f:Figure5}
639: \begin{figure} [!ht]
640: \centering\includegraphics[width=\columnwidth]{images/Figure6.pdf}
}
642: \label{f:Figure6}
676: \begin{figure} [!ht]
677: \centering\includegraphics[width=\columnwidth]{images/Figure7.pdf}
}
679: \label{f:Figure7}
716: \begin{figure} [!ht]
717: \centering\includegraphics[width=\columnwidth]{images/Figure8.pdf}
}
721: \label{f:Figure8}
747: \begin{figure} [!ht]
748: \centering\includegraphics[width=\columnwidth]{images/Figure9.pdf}
}
750: \label{f:Figure9}
```

```
768: \begin{figure} [!ht]
769: \centering\includegraphics[width=\columnwidth]{images/Figure10.pdf}
    f}
772: \label{f:Figure10}
797: \begin{figure} [!ht]
798: \centering\includegraphics[width=\columnwidth]{images/Figure11.pdf}
    f}
801: \label{f:Figure11}
```

figures 11

tables 2

includegraphics 11

labels 13

refs 0

floats 13

True : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

False : check if all figures are referred to: (refs >= labels)

Label/ref check

91: abusing prescribed opioid medication who also used heroin, shown  
in Figure 1.

335: 386 females). Table 1 shows the raw counts of individual  
substance use by age

394: Table 2 shows the frequency of individuals reporting that they  
had experienced

423: Figure 1 shows the proportion of individuals who reported  
misusing prescription

425: Figure 1 shows the majority of respondents (89 percent) stated  
they had never

428: Figure 1 shows that, of those individuals who reported using  
heroin, the

443: Figure 2 shows the aggregated measure of Opioid Pain Reliever  
misuse and abuse

471: Figure 3 shows the pairplots of demographic features including  
mental health

537: few features. Figure 4 plots the coefficients of logistic  
regression classifier

543: accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows  
that the

574: Figure 5 shows a partial view of the decision tree classifier of  
heroin use

627: feature importance for a single tree. Figure 6 shows the feature  
importance

667: or by reducing the learning rate. Figure 7 shows that the feature  
importance  
705: Figure 8 plots the coefficients of logistic regression classifier  
for  
707: Figure 8, the features with the highest coefficient values were  
Treatment  
730: of the pruned tree was 0.902 and test set accuracy was 0.902.  
Figure 9 shows  
733: notebook) \cite{classifyPRL}. As Figure 9 shows, the decision  
tree classifier  
758: 0.896, which suggests that the model overfit the data. Figure 10  
shows the  
758: 0.896, which suggests that the model overfit the data. Figure 10  
shows the  
760: PRL Misuse. As Figure 10 shows, several features were identified  
as important  
760: PRL Misuse. As Figure 10 shows, several features were identified  
as important  
786: prescription opioid abuse. Figure 11 shows the feature importance  
for the  
786: prescription opioid abuse. Figure 11 shows the feature importance  
for the  
787: gradient boosting classifier tree. As Figure 11 shows, several  
features were  
787: gradient boosting classifier tree. As Figure 11 shows, several  
features were  
820: used heroin than had not (see Figure 1), which partially supports  
the  
passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

WARNING: algorithm and below may be used improperly

128: classification algorithms are considered below.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "granovetter73"
Warning--I didn't find a database entry for "watts98"
Warning--page numbers missing in both pages and numpages fields in herland14
Warning--no number and no volume in johnson11
Warning--page numbers missing in both pages and numpages fields in johnson11
(There were 5 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
-----
```

```
passed: True
```

# IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)

Ashok Reddy Singam

Indiana University

711 N Park Ave

Bloomington, Indiana 47408

asingam@iu.edu

Anil Ravi

Indiana University

711 N Park Ave

Bloomington, Indiana 47408

anilravi@iu.edu

## ABSTRACT

The predictive health management (PHM) is an enabling discipline consisting of technologies and methods to assess the reliability of a product in its actual life cycle conditions to determine the advent of failure and mitigate system risk. The PHM system will monitor environmental, operational, and performance related characteristics of the product and gathered data analyzed to assess product health and predict remaining life.

In this application, the industrial rotating equipment such as compressors, vacuum blowers, pumps, and valves etc. are considered to monitor and analyze their operational behavior. The product critical operational parameter data such as vibration, temperature, and load current will be collected from field sensors and analyzed to predict the failure using kNN machine learning classification algorithms. The data will be collected from the field using wireless sensors and stored on the cloud based AWS database server. The product data will be analyzed and made available to all stakeholders to take appropriate preventive actions via web/mobile applications.

## KEYWORDS

i523, HID333, HID337, KNN, IoT, Big Data, Analytics

## 1 INTRODUCTION

The PHM technology can be put within a broader business context by relating it to the Product-Service System (PSS) business model. PSS can be defined as an integrated combination of products and services where the emphasis is put on the “sale of use” rather than the “sale of product” [2]. Central to this new business model is a shift from selling a product, and its related spare parts as required, to selling a solution that supports customer needs in the form of a service delivering a fully maintained and useable product [2]. As shown in Figure 1, There are several wireless technologies such as 802.11, cellular, and short distance wireless protocols can be used to collect and send data to the centralized servers. Also, data can be stored in cloud based technologies such as AWS, Microsoft Azure, IBM and Google etc. for processing.

[Figure 1 about here.]

**Problem Statement:** in the manufacturing operations, automotive and other process industries rotating equipment such as pumps, valves, compressors, and blowers are commonly used equipment for various purposes. These equipment are severely suffered from wear and tear, bearing degradation, shaft misalignment, corrosion, and other mechanical breakdowns. Due to the limitations of wireless enabled sensors based data acquisition it was very difficult to collect this data in the past. Also, due to real-time nature of the data

acquisition, it was a huge challenge to store the data locally and process the information for applying machine learning algorithms. All these technological and infrastructure limitations caused industrial equipment health monitoring had become one of the sector businesses are losing the money due to operation shutdowns and unplanned maintenance etc.

**Solution Approach:** with the wireless sensors and cloud based server technologies, it has become possible to deploy hundreds of sensors in the manufacturing plant and collect the data and store with minimal costs. Once the data is stored on the servers with high computing power, machine learning algorithms can be used to process the sensor data to predict the equipment failures with reasonable accuracy. This approach has been named as predictive or prognostics health management of the equipment which is widely available in the recent times due to the availability of technological infrastructure.

The PHM generally combines sensing, collecting, storing and analyzing of environmental, operational, and performance related parameters to assess the health of a product and predict remaining useful life. Assessing the health of a product provides information that can be used to meet several critical goals [1]:

- Providing advance warning of failures
- Minimizing unscheduled maintenance, extending maintenance cycles, and maintaining effectiveness through timely repair actions
- Reducing the life cycle cost of equipment by decreasing inspection costs, downtime, and inventory
- Improving qualification and assisting in the design and logistical support of fielded and future systems

The PHM is not a new concept, however, with the advent of sensors, machine learning algorithms, and computing capacity of the servers it has become more prevalent in the recent days. In this application, an attempt has been made to prove the concept of simple PHM implementation and use in real world applications. The application can be re-architected to address more complex products/systems with considerations of scalability, performance, cost and reliability. The limitations of the current application are described in the end of this report.

The parameter monitoring and the analysis of acquired data using prognostic models are fundamental steps for the PHM methods. The sensors are the essential devices used to monitor parameters and obtain long-term accurate information to provide anomaly detection, fault isolation, and rapid failure prediction [1].

Firstly, PHM requires monitoring a large number of product parameters to evaluate the health of a product. Depending on the

complexity of the monitored product, it is possible to monitor thousands of parameters in the entire life cycle of the product to provide the information required by PHM. These parameters include operational and environmental loads as well as the performance conditions of the product, for example, temperature, vibration, shock, pressure, acoustic levels, strain, stress, voltage, current, humidity levels, contaminant concentration, usage frequency, usage severity, usage time, power, and heat dissipation. In each case, a variety of monitoring features such as magnitude, variation, peak level, and rate of change may be required in order to obtain characteristics of parameters [1].

In this application, commonly used equipment in industrial and automobile operations such as air compressors, vacuum blowers, and smart valves are considered for analysis. The critical operational parameters of these products will be collected using applicable sensors from the field and fed to a database at regular intervals.

In general design, the frequency of data collection and storage depends on the number of parameters to be analyzed, cost of the system and operational behavior of the equipment. For this application, since products with rotating parts are considered, the critical parameters that would define the health of the equipment are: input or load current, internal ambient temperature, and vibration of the equipment.

The PHM application design process is shown in Figure 2, which describes various steps of the processes involved. For the implementation of this project, the sensor generated data is simulated using SQL scripts due to development time constraints. However, a detailed step-by-step approach is provided if we need to plug-in the sensor modules in to the application.

[Figure 2 about here.]

**Data Acquisition Stage:** It is required to have a description of a machine behaving normally that can be used for early detection of anomalies. This calls for a proper characterization of machine health. As part of this process, various methods are identified to extract health information from vibration measurements and investigate strengths and weaknesses of these methods as health descriptors. This stage will be the core part of PHM application where vibration data were experimentally obtained from a compressor using triaxial accelerometer to collect transverse, longitudinal and vertical axes vibration signals. For the experimental data collection, ACC301A triaxial accelerometer and National Instruments data acquisition system was used. A total of 8 parameters

- (1) Input Current
- (2) Input Voltage
- (3) Internal ambient temperature
- (4) External ambient temperature
- (5) Transverse vibration
- (6) Longitudinal vibration
- (7) Vertical axis vibration
- (8) Acquisition time

were captured at one second rate, which generated about 65000 records. This data has been analyzed for identifying the feature classification.

**Pre Processing stage:** During this stage collected data will be filtered and processed for accuracy in order to adapt them to subsequent feature extraction stage. In this application, all the

pre-processing has been done manually to validate the accuracy of the data based on the system conditions. Since spectral analysis of vibration signals are not done ( one of the limitations for this application, captured in the end), the data generated from the compressor is considered as the primary frequency of the equipment (which is isolated from the rest of the attachments).

**Feature extraction and selection stage:** during this stage domain specific vibration spectral analysis has been performed but only considered time-domain behavior for various system operational conditions such as increased load, modified input voltage, and modified external ambient temperature etc. Based on the response of the machine vibration to various external conditions were noted down. This data is used to identify the following feature vectors.

- NORMAL OPERATION AT 30 DEG CENTIGRADE
- OVER CURRENT FAULT OPERATION
- OVER TEMPERATURE FAULT OPERATION
- INPUT OVER VOLTAGE FAULT OPERATION
- ABNORMAL OPERATION AT 30 DEG CENTIGRADE
- BEARING DEGRADATION OPERATION

**kNN classification stage:** this stage is core part of the PHM application, which will predict the unknown test data to be classified in to a known label based on the training data set using nearest neighbor algorithm.

**Classifier performance evaluation stage:** this stage will be used to evaluate the classifier accuracy of prediction. In this application, k-fold cross-validation method has been used to perform the evaluation.

The data is generated and made available in Oracle database on AWS cloud to perform analysis. The application developed in this project will consist of the following components:

- Sensor Data Generator
- Machine Learning Algorithm
- Big Data and IoT
- PHM Dashboard
- Decision Alerts
- Application Script

The following sections will describe the architectural and design aspects of the PHM system implementation in detail.

## 2 PROGNOSTICS MODEL EVALUATION

The prediction is typically performed only after the *health* of the component or system deteriorates beyond a certain threshold [7]. In this application, faults and failures are identified in the training data set. The faults identified are: Over current fault, over temperature fault and over voltage fault. If over current fault is occurred, the equipment will tend to draw higher current than nominal values which if continued further several times eventually leads to a permanent failure of the equipment. In this application, when motor bearing starts degrading, the first observation will be over current followed by over temperature conditions. Often times, that threshold is tripped because a fault occurs. A fault is a state of a component or system that deviates from the normal state such that the integrity of the component is outside of its required specification. A fault does not necessarily imply that the overall system does not operate anymore; however, the damage that characterizes the fault often grows under the influence of operations to a failure. The

latter is the state at which the component or system does not meet its desired function anymore. It is the task of prognostics to estimate the time that it takes from the current time to the failed state, conditional on anticipated future usage. This would give operators access to information that has significant implications on system safety or cost of operations. Where safety is impacted, the ability to predict failure allows operators to take action that preserves the assets either through rescue operation or through remedial action that avert failure altogether. Where minimizing cost of operations is the primary objective, predictive information allows operators to avert secondary damage, or to perform maintenance in the most cost-effective fashion. Often times, there is a mix of objectives that need to be optimized together, sometimes weighted by different preferences [7].

As emphasized above, predictive models evaluation needs to take domain specificities into account. Such specificities cover two aspects: capability of failure prediction and TTF estimation. From the point of view of TTF, it is desirable that a predictive model can generate alerts in a *targeted* time window prior to a failure. A model that predicts a failure too early leads to non-optimal component use which will impact the reliability or availability of the system [9].

[Figure 3 about here.]

As shown in the Figure 3, the time to failure prediction will be estimated based on the classified result data set and alert the stakeholders to take relevant actions. The target alert zone will be identified based on the abnormal behavior of the equipment over the period.

### 3 APPLICATION DESIGN ANALYSIS

The PHM application in this project considered to use rotating equipment temperature, load current and vibration data for analyzing and predicting the future operational behavior. Vibration signals from rotating components are usually analyzed in the frequency domain, because significant peaks in the signal spectrum appear at frequencies that are related to the rotation frequency of the component. In this application, only time domain parameters with peak vibration magnitudes irrespective of the frequency component. The training data set consists of normal, abnormal, and fault conditions vibration patterns describes the system characteristics from which its status can be estimated. The PHM application for industrial equipment machine failure detection problem directly correlates to the pattern classification problem. From the vibration data collected, each accelerometer will output values of X, Y, and Z data then using a KNN we can similarly identify which vibration parameter(s) determines problems in our machines, or *likely to experience failure*. The typical defects or failures that can be detected are: machine imbalance, shaft misalignment, pumps cavitation, structural and rotating looseness, early stage bearing wear, gear teeth problems, and other high-frequency defects.

This application used *Sensor Data Gen SQL* script module to generate the sensor data and store in Oracle database on AWS. This is the critical module as we have not used the real data collection from the field. However, the sensor hardware and necessary environment to generate the data is identified and experimented to

work with. A brief description about the hardware is provided in the end of this report.

The PHM application is designed such that the fundamental concepts can be verified to open a discussion on limitations, performance, scalability, ROI and reliability of the system.

The following sections describe the application design components with necessary implementation details:

#### 3.1 Sensor Data Generator

The SQL data generator script is designed to generate training data as well test data for this application with following eleven parameters: Acquisition time, equipment name, part number, serial number, internal ambient temperature, external ambient temperature, input voltage, input current, and vibration data for x, y, and z axes. The following database design architecture followed for Sensor Data Gen module:

- Sensor Data Generator PL SQL Objects
  - Tables
    - \* SENSOR TRAIN DATA for storing training data
    - \* SENSOR TEST DATA for storing testing data
  - Views
    - \* SENSOR TRAIN DATA VIEW: Created View on top of SENSOR TRAIN DATA with logic to translate string label data into numbers
    - \* SENSOR TEST DATA VIEW Created View on top of SENSOR TEST DATA with logic to translate string label data into numbers
  - Packages BIG DATA 503 PRJ PKG
    - \* Generate Test Set: Pl/Sql procedure to insert sensor test data into SENSOR TRAIN DATA table
    - \* Generate Train Set: Pl/Sql procedure to insert sensor train data into SENSOR TRAIN DATA table
    - \* Delete Data Set: Pl/Sql procedure to delete all training and test data.
    - \* Update Test Data Labels: Pl/Sql procedure to update SENSOR TRAIN DATA table with KNN algorithm predicted label values

#### 3.2 Machine Learning Algorithm

3.2.1 *Classifier evaluation.* Typical classifier evaluation methods include ROC Curves, Reject Curves, Precision-Recall Curves, and Statistical Tests. The statistical tests consists of following methods to perform evaluation:

- Estimating the error rate of a classifier
- Comparing two classifiers
- Estimating the error rate of a learning algorithm
- Comparing two algorithms

Out of the listed statistical tests, the error rate estimation method is used in this application to evaluate the performance. An experimental data is used to estimate the error rate or accuracy of various classifiers. Then a comparison has been made to choose the classifier to use in the application.

The following list of performance for various classifiers is observed during the accuracy calculation. The same set of training

data has been used for all the classifiers, which has resulted the following performance. All values are mentioned in percents between 0 to 1, 1 means 100 percent accuracy.

- LogisticRegression: 0.963636
- KNN: 0.981818
- DecisionTreeClassifier: 0.964394
- SVM: 0.972727

Based on the performance as shown in Figure 4, kNN has been selected to use for this application.

[Figure 4 about here.]

**3.2.2 *k Nearest Neighbor - kNN*.** In this application, neighbors-based classification is chosen to classify the unknown instance to the known trained labels [10]. Neighbors-based classification does not attempt to construct a general internal model, but simply stores instances of the training data.

Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consisting of training observations  $(x, y)$  and would like to capture the relationship between  $x$  and  $y$ . More formally, our goal is to learn a function

$$h : X - \rightarrow Y$$

so that given an unseen observations  $x$ ,  $h(x)$  can confidently predict the corresponding output  $y$ .

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the  $K$  most similar instances to a given unseen observation. The number of neighbors for  $k$ -nearest neighbors ( $k$ ) can be any value less than the number of rows from dataset. Looking at only a few neighbors makes the algorithm perform better but the less similar the neighbors, the worse the prediction will be. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \dots + (x_n - x'_n)^2}$$

But other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance. An alternate way of understanding KNN is by thinking about it as calculating a decision boundary (i.e. boundaries for more than 2 classes) which is then used to classify new points.

Another characteristic of KNN is it is instance based learning algorithm. Means it doesn't explicitly learn a model. Instead, it chooses to memorize the training instances which are subsequently used as knowledge for the prediction phase. It is also means the algorithm does not build a model until the time that a prediction is required. It is also lazy learning because it only does work at the last second. This has the benefit of only including data relevant to the unseen data, called a localized model. A disadvantage with lazy model is it can be computationally expensive to repeat the same or similar searches over larger training datasets.

In the application design, sci-kit open source python libraries are used for implementing the kNN algorithms. Scikit is built on NumPy, SciPy, and matplotlib. The k-neighbors classification in KNeighborsClassifier is the more commonly used of the two techniques. The optimal choice of the value  $k$  is highly data-dependent: in general a larger  $k$  suppresses the effects of noise, but makes the classification boundaries less distinct.

The sklearn.neighbors.KNeighborsClassifier class has the following methods, which are used in the application design [8]:

- fit: Fit the model using  $X$  as training data and  $y$  as target values.
- get params: Fit the model using  $X$  as training data and  $y$  as target values.
- kneighbors: Finds the  $K$ -neighbors of a point.
- kneighbors graph: Computes the (weighted) graph of  $k$ -Neighbors for points in  $X$ .
- predict: Predict the class labels for the provided data.
- predict\_proba: Return probability estimates for the test data  $X$ .
- score: Returns the mean accuracy on the given test data and labels.
- set params: Set the parameters of this estimator.

**3.2.3 *K-fold cross-validation*.** To estimate the test error in the model, a cross-validation approach followed in which a subset of the training set will be holding out from the fitting process. This subset, called the validation set, can be used to select the appropriate level of flexibility of our algorithm. There are different validation approaches that are used in practice, and we will be exploring one of the more popular ones called **k-fold cross validation**. The k-fold cross validation (the  $k$  is totally unrelated to  $K$ ) involves randomly dividing the training set into  $k$  groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining  $k - 1$  folds. The misclassification rate is then computed on the observations in the held-out fold. This procedure is repeated  $k$  times; each time, a different group of observations is treated as a validation set. This process results in  $k$  estimates of the test error which are then averaged out.

In this application, an average k-fold cross validation accuracy of 0.99 percent achieved, which is explained in the appendix section of the report. Figure 5 shows Classification and Confusion report output obtained from the KNN model we used for this project.

[Figure 5 about here.]

### 3.3 Big Data and IoT

In PHM systems big data is characterized by one or more 3Vs: volume, velocity and variety due to streaming of real-time IoT sensors. Most of the IoT systems present challenges in combinations of velocity and volume. The important feature of the IoT application is that by observing the behavior of "many things" it will be possible to gain important insights, optimize processes, etc. This requires storing all the events (velocity and volume challenge) to run analytical queries over the stored events and perform analytics (data mining and machine learning) over the data to gain insights. In general PHM applications, data will be collected through field sensors at specific rate which accounts for large amount of data per day in the order of multi-million records. This data will be stored

in any NOSQL or RDBMS based database for storage and processing. Since the big data infrastructure is much reliable and available widely from multiple vendors, it would help to build complex PHM systems with large number of feature vectors for classification.

In this application, for the demonstration of the concept, the real vibration data from the compressor equipment has been collected via accelerometer sensors. This vibration data has been analyzed in time domain and established the labels based on the compressor design performance parameters. Later, this data analysis is used to design a SQL script for generating training and test data sets. However, the real-time PHM system will have continuous streaming of data coming from hundreds of devices at faster rates (in the order of milliseconds to tens of seconds). This data needs to be captured by reliable and scalable platforms such as AWS IoT or similar and use the machine learning algorithms to classify the unknown data.

### 3.4 PHM Dashboard

Once all the test data set has been classified in to appropriate labels, the prediction of the failure can be performed based on the trending of the equipment behavior over the period. In order to understand the equipment performance insight, following queries will be used on the classified data:

- Faults Reported by Equipment Part Number
- Faults Reported by Serial Number
- Abnormal Behavior by Equipment Part Number
- Abnormal Behavior by Serial Number over the period range

There can be more application specific information obtained from classified data set to take various decisions. Figure 6, Figure 7 and Figure 8 show various PHM data analytics for this project. Figure 6 displays all the serial numbers of equipment 1 with bearing degradation problems. The X axis gives serial numbers while the Y axis gives number of occurrences of bearing degradation for that particular serial number. Similarly Figure 7 and Figure 8 give the details of over temperature and over current faults of various serial numbers.

As part of data visualization, result data file is queried based on the analytics metrics interested. The python matplotlib package has been used to draw the charts as needed for showing the analytics. In real world application, a more sophisticated business intelligence tools such as Tableau, Microsoft BI, and Amazon Quick Sight can be used to show the PHM dashboards. These dashboards are targeted for business users so that they will be able to customize the views, add filters and drill down in to specific information as needed.

Sample screen shots for the following scenarios are included from python code output:

- Bearing Degraded Serial numbers for Equipment Part Number1: Figure 6
- Over Temperature Fault Serial numbers for Equipment Part Number1: Figure 7
- Over Current Fault Serial numbers for Equipment Part Number1: Figure 8

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

### 3.5 Decision Alerts

Once the results data set has been generated by the prediction algorithm, and then based on the analytics queries, PHM system can send out the alert messages to appropriate stakeholders. The typical messages include the following s minimum:

- SN 10002: Faulted X times on over temperature in last Y days, needs maintenance to clean the filter
- SN 10005: Consistently indicating bearing degradation from last X days, needs lubrication maintenance
- SN 10009: Consistently drawing over current from last X days, needs mechanical load maintenance

In this application all the unseen test data is classified and labeled in the result data file. However, in real world application, along with the dashboards a comprehensive alerting capabilities can be built. The application checks for out of range alert conditions on selected incoming report parameters, looking for warning or alarm conditions that are higher or lower than expected under normal operating conditions.

### 3.6 Application Code Development

Code required for this project is divided into two categories

- Python Coding: We used **Anacoda Navigator ver 1.6.9** installation on windows7 laptop which includes Jupyter notebook application for python coding [3]. **Anacoda Navigator** also supports multiple installation and management of python environments using gui interface. The location of the Jupyter notebook file we developer for this project is mentioned in appendix b.
- PL/SQL Coding: This application specific training/test sensor data has been created/generated on AWS cloud database using pl/sql coding which will be accessed during the run time by python notebook. We used **Orcale Sql Developer** for pl/sql coding [4].

### 3.7 Python - Oracle Interface

For this project we used python library called **cxOracle** to enable access to Oracle Database [5]. It can be installed easily using **pip** and it supports both Python 2 and 3. This library supports:

- SQL and PL/SQL Execution. The underlying Oracle Client libraries have significant optimizations including compressed fetch, pre-fetching, client and server result set caching, and statement caching with auto-tuning.
- Extensive Oracle data type support, including large object support like CLOB and BLOB)
- Batch operations for efficient INSERT and UPDATEs

In the following scenarios we used **cxOracle** libraries:

- To read sensor training data from Oracle database
- To read sensor test data from Oracle database
- After classification of labels, to update test data with classified labels

## 4 APPLICATION LIMITATIONS

The PHM application developed in this project has several limitations. Typically, PHM applications suffer from the prediction accuracy rate which influences the ability to take decisions that

will have broader impact on the business operations and financial aspects. However, with the advanced machine learning classifiers and model evaluation methods this can be addressed to achieve reasonable confidence. Following are some of the limitations of this application, which can be addressed and improved in large real-time PHM systems.

**Data acquisition hardware:** in this application the data is not collected from real-time sensors for voltage, current, temperature and vibration data. There will be inherent accuracy in the raw data generated by SQL script. However, a sample vibration dataset has been collected from the field, which is used as basis to generate the simulated data set.

**Feature extraction analysis:** the equipment performance parameters of interest need to be down selected from large set of incoming parameter data.

When analyzing vibration data in the time domain only few parameters are available in quantifying the strength of a vibration profile: amplitude, peak-to-peak value, and RMS. The amplitude is valuable for shock events but it does not take into account the time duration and thus the energy in the event. The same is true for peak-to-peak with the added benefit of providing the maximum excursion of the wave, useful when looking at displacement information, specifically clearances. The RMS value is generally the most useful because it is directly related to the energy content of the vibration profile and thus the destructive capability of the vibration.

This requires in-depth domain specific analysis, in this case a detailed mathematical modeling of vibration spectral analysis to precisely select the features and corresponding behavior patterns. Such analytical data should be used for training data feature set. In this application, a primitive approach of time-domain analysis of vibration magnitudes used for determining features. However, in real application these features need to be mathematically analyzed to identify the features that represent the system behavior as close as possible.

**Model accuracy and scoring :** the kNN algorithm used in this application validated using k-cross fold cross-validation. There are several other model evaluation and scoring methods such as accuracy (or error rate), True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), sensitivity etc. These metrics provide a simple and effective way to measure the performance of a classifier. This application can be further improved by applying more performance measurement methods to increase the effectiveness of the algorithm design.

**Scalability:** the application designed in this project is very primitive to understand the basic concepts of PHM and kNN classifier implementation. This application cannot be used for PHM application in business use. To implement real world PHM application, a more comprehensive design needed by considering modularity, service oriented architecture, large number of sensors integration, big data and analytics integration etc.

## 5 RECOMMENDATIONS

**Generality:** since each rotating equipment vibrates in a different manner, a monitoring method needs to be retrained for each machine. The training on several repeated measurements on several

similar equipment in several operating modes may allow for a more general monitoring method.

**Feature extraction and dimensionality:** In this application it has been assumed that a proper feature (selection) has been chosen, such that the feature dimensionality is not too high. If the data lies in a subspace, application of an initial dimensionality reduction may be a good idea. It is highly recommended to perform spectral analysis on vibration data and identify various fault frequencies and their sources. This would help to extract the optimized feature vectors for the given application followed by selecting the more relevant ones.

**Model evaluation:** classifier accuracy and effectiveness will be varied based on the test data set. It is highly recommended to evaluate multiple models with appropriate test data to choose the best classifier for the given application.

**Domain specific modeling:** It is highly recommended to perform more and more domain-oriented feature vector analysis to meet the needs of predictive model evaluation for PHM applications. Domain-oriented approaches helpful and useful in evaluating classifier for applications. Generic evaluation methods could help developers in investigating overall performance of a model from the statistical viewpoint at the initial stage of model development. Domain-oriented approaches should be further used to evaluate the usefulness and business value [9].

## 6 CONCLUSION

In this project, the problem statement around industrial rotating equipment maintenance is described and solution principle to address the same using PHM concept is defined, experimented and results are discussed. Since this application is developed to prove the only concept but not the complete solution a section with limitations and recommendations for real world system development is described. Overall, PHM application with kNN classifier algorithm and cross validation accuracy of 0.99 percent has been implemented, verified and results are analyzed for business decisions.

## REFERENCES

- [1] Shunfeng Cheng, Michael H. Azarian, and Michael G. Pecht. 2010. Sensor Systems for Prognostics and Health Management. (2010), 24 pages. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3247731/pdf/sensors-10-05774.pdf>
- [2] Tonci Grubic, Ian Jennions, and Tim Baines. 2009. The Interaction of PSS and PHM - a mutual benefit case. (2009), 10 pages. [https://www.phmsociety.org/sites/phmsociety.org/files/phm\\_submission/2009/phmc\\_09\\_49.pdf](https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2009/phmc_09_49.pdf)
- [3] Anaconda Inc. 2017. Anaconda Python Data Science platform. (2017). <https://www.anaconda.com/what-is-anaconda/>
- [4] Oracle. 2017. Oracle SQL Developer. (2017). <http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index.html>
- [5] Oracle OTN. 2005. Using Python With Oracle Database 11g. (2005). <http://www.oracle.com/technetwork/articles/dsl/python-091105.html>
- [6] RuizGonzalez Ruben, GomezGil Jaime, GomezGil Francisco Javier, and Martinez Victor. 2014. An SVM Based Classifier for Estimating the State of Various Rotating Components in Agro Industrial Machinery with a Vibration Signal Acquired from a Single Point on the Machine Chassis. *Sensors* 14, 11 (2014), 20713–20735. <https://doi.org/10.3390/s141120713>
- [7] Abhinav Saxena, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. 2009. Sensor Systems for Prognostics and Health Management. (2009), 16 pages. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100023445.pdf>
- [8] scikit learn.org. 2017. sklearn.neighbors.KNeighborsClassifier. (2017). <http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
- [9] Chunsheng Yang, Yanni Zou, Jie Liu, and Kyle R Mulligan. 2014. Predictive Model Evaluation for PHM. (2014), 11 pages. [https://www.phmsociety.org/sites/phmsociety.org/files/phm\\_submission/2014/ijphm.14.019.pdf](https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2014/ijphm.14.019.pdf)

- [10] Kevin Zakka. 2016. A Complete Guide to K-Nearest-Neighbors with Applications in Python and R. (2016). <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

## A WORK BREAKDOWN

### A.1 HID 333:Anil Ravi

- Identified Project topic.
- Created architecture of the application.
- Ran experimental test to collect vibration data
- Extracted and analyzed feature vectors
- Studied, designed and reviewed kNN algorithm
- Created draft project report
- Reviewed the draft project report.

### A.2 HID 337:Ashok Reddy Singam

- Implemented sensor data generation SQL script.
- Implemented kNN algorithm in Python
- Implemented k-fold cross validation design
- Created data analytics charts
- Reviewed the draft project report.

## B CODE REFERENCE

All code, notebooks and files for this project can be found in the github repository: <https://github.com/bigdata-i523/hid337/blob/master/project/jupyter>

## LIST OF FIGURES

1	System Architecture	9
2	PHM Design Process [6]	9
3	Time relation between alert time and failure time [9]	10
4	Algorithm performance	10
5	KNN Classification and Confusion matrix report	11
6	Bearing Degradation	11
7	Over Temperature Fault	12
8	Over current Fault	13

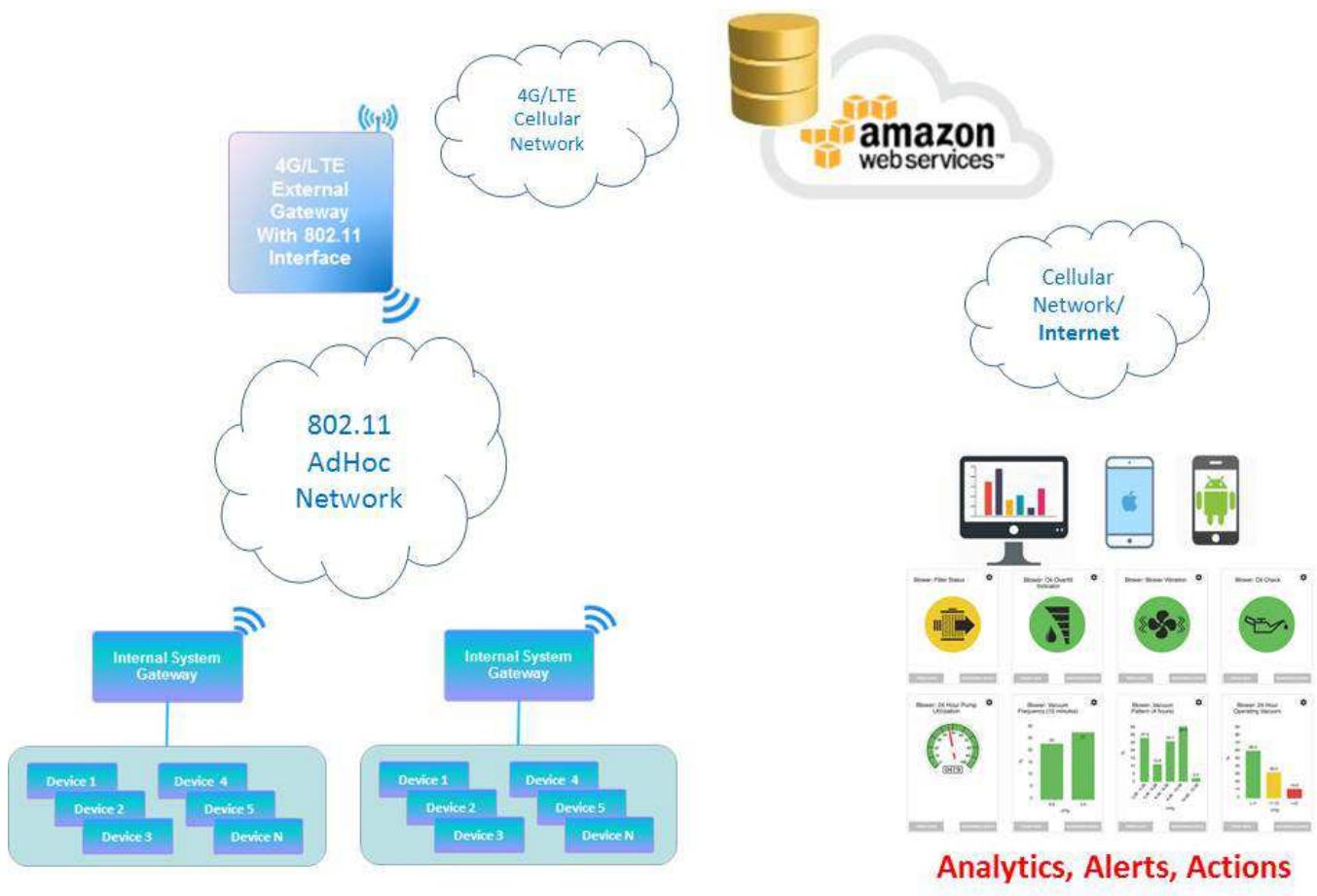


Figure 1: System Architecture

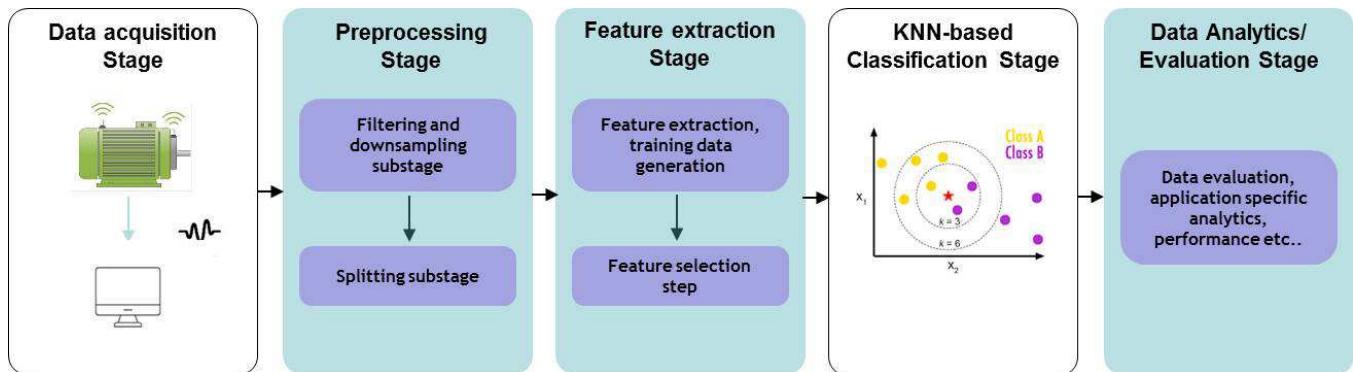


Figure 2: PHM Design Process [6]

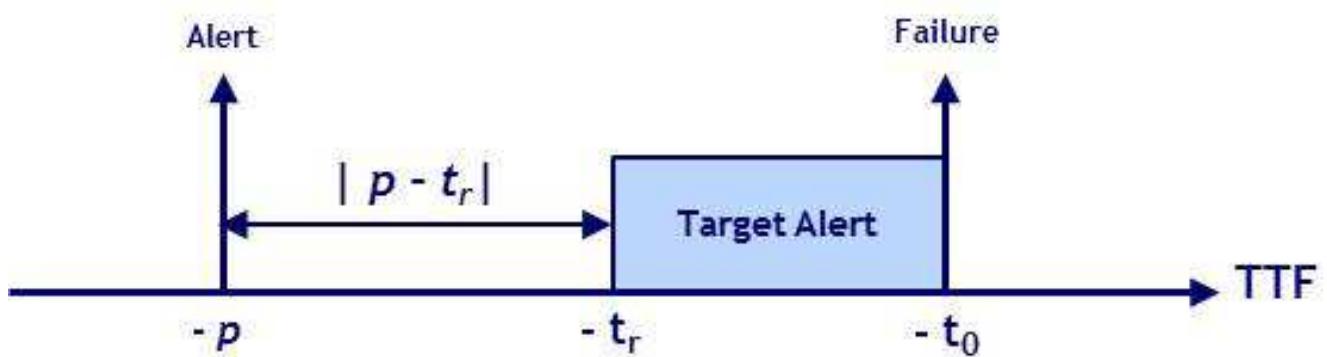


Figure 3: Time relation between alert time and failure time [9]

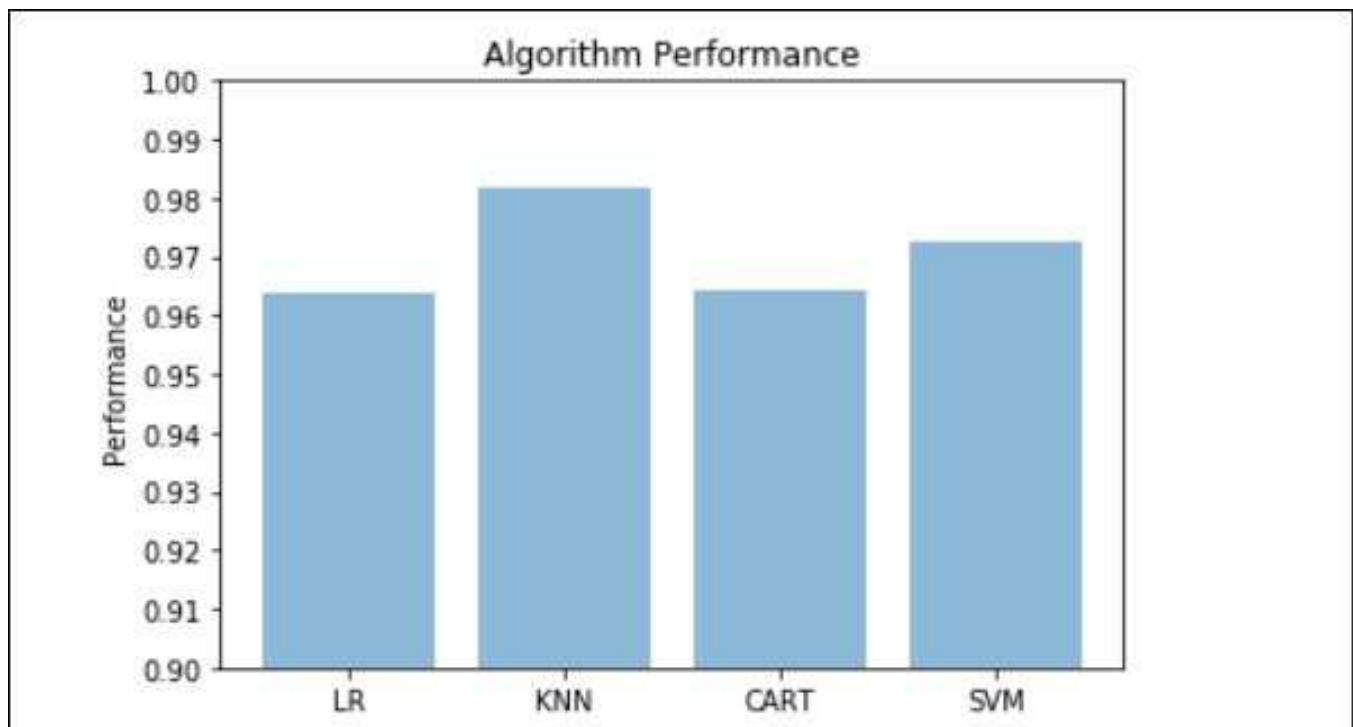


Figure 4: Algorithm performance

	precision	recall	f1-score	support
ABNORMAL_OP_30_DEG_C	1.00	1.00	1.00	997
BEARING_DEGRADE_OP	1.00	1.00	1.00	100
INPUT_OVER_VOLT_FAULT_OP	0.99	0.96	0.98	1059
NORMAL_OP_30_DEG_C	0.96	0.99	0.97	931
OVER_CURRENTFAULT_OP	1.00	1.00	1.00	1005
OVER_TEMP_FAULT_OP	1.00	1.00	1.00	1130
avg / total	0.99	0.99	0.99	5222
[[ 997 0 0 0 0]				
[ 0 100 0 0 0]				
[ 0 0 1018 41 0]				
[ 0 0 10 921 0]				
[ 0 0 0 0 1005]				
[ 0 0 0 0 1130]]				

Figure 5: KNN Classification and Confusion matrix report

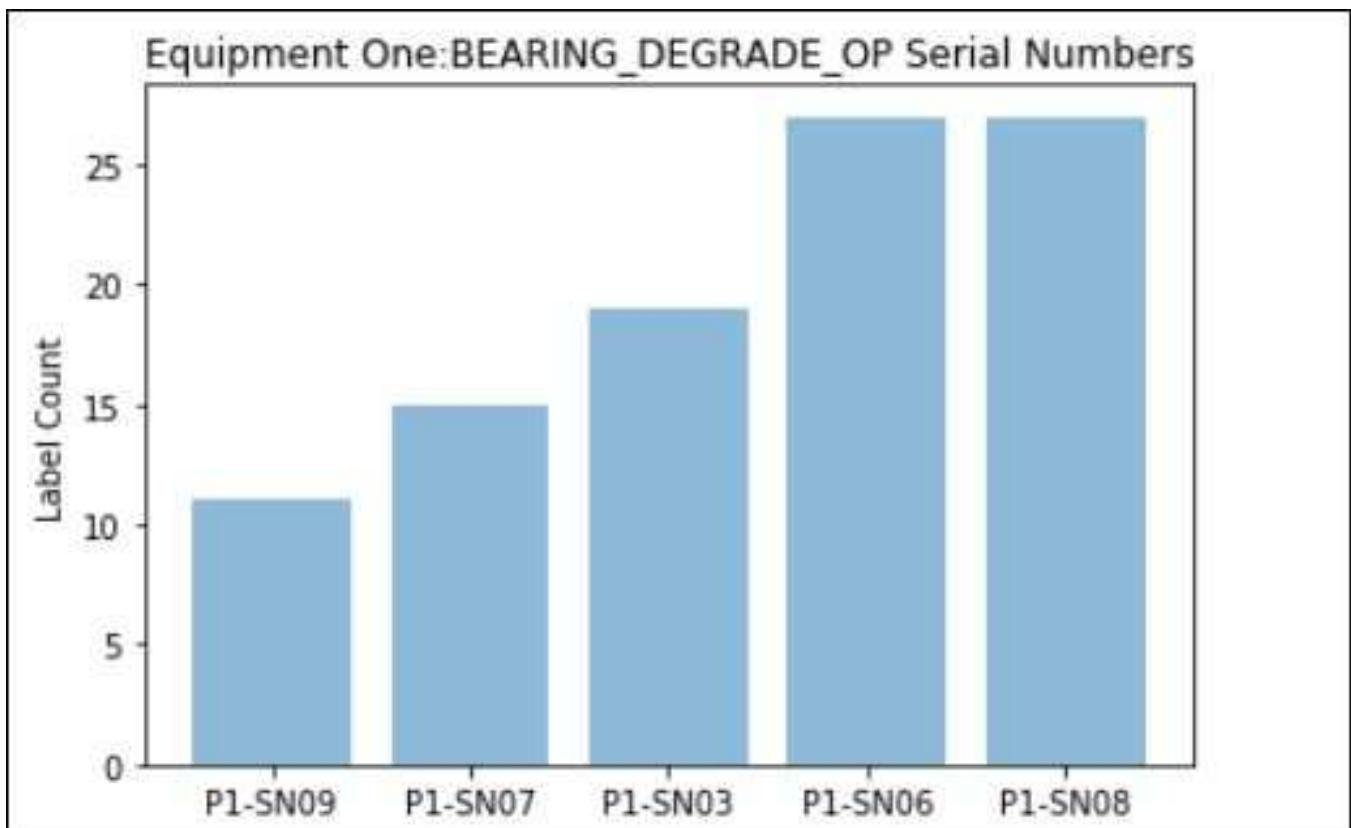


Figure 6: Bearing Degradation

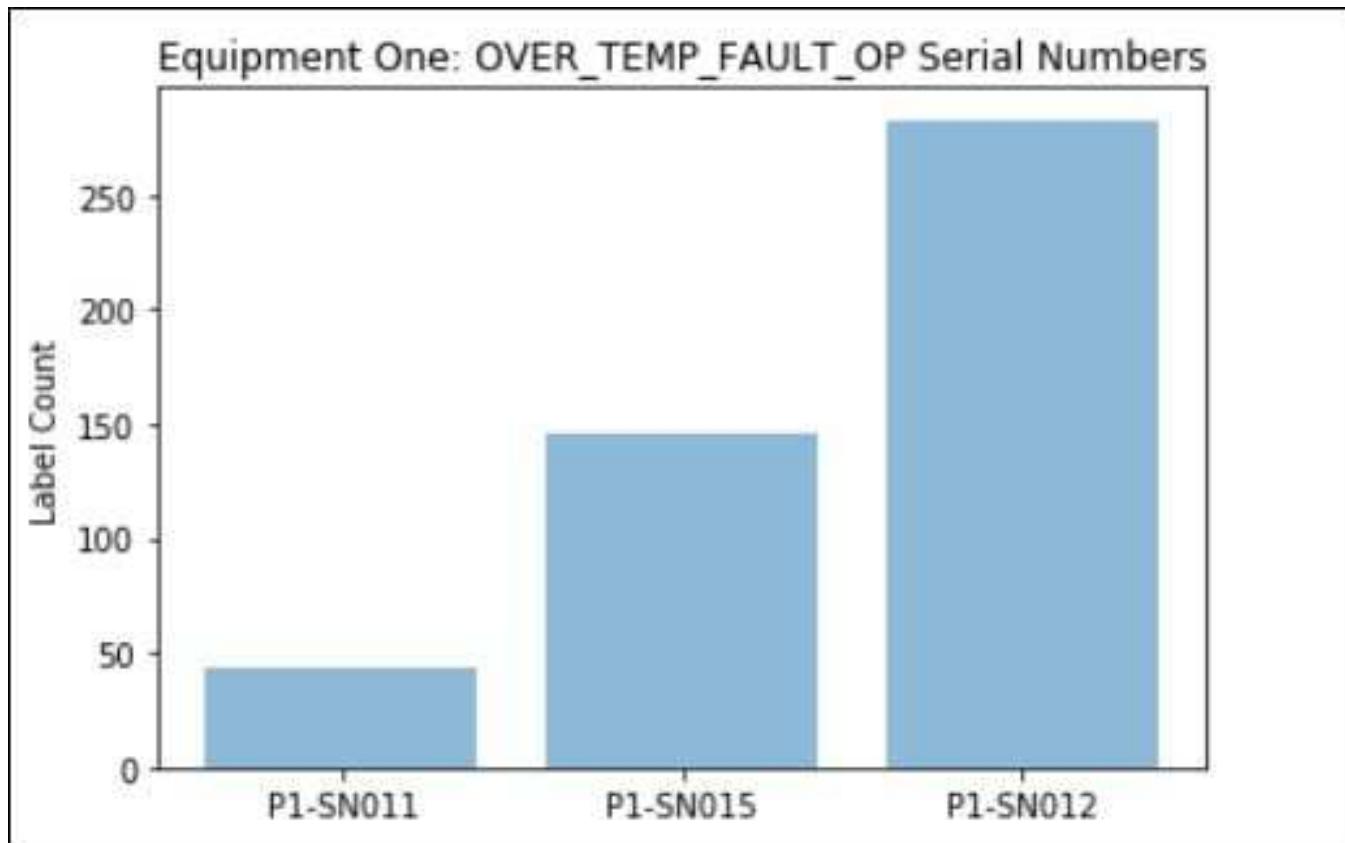


Figure 7: Over Temperature Fault

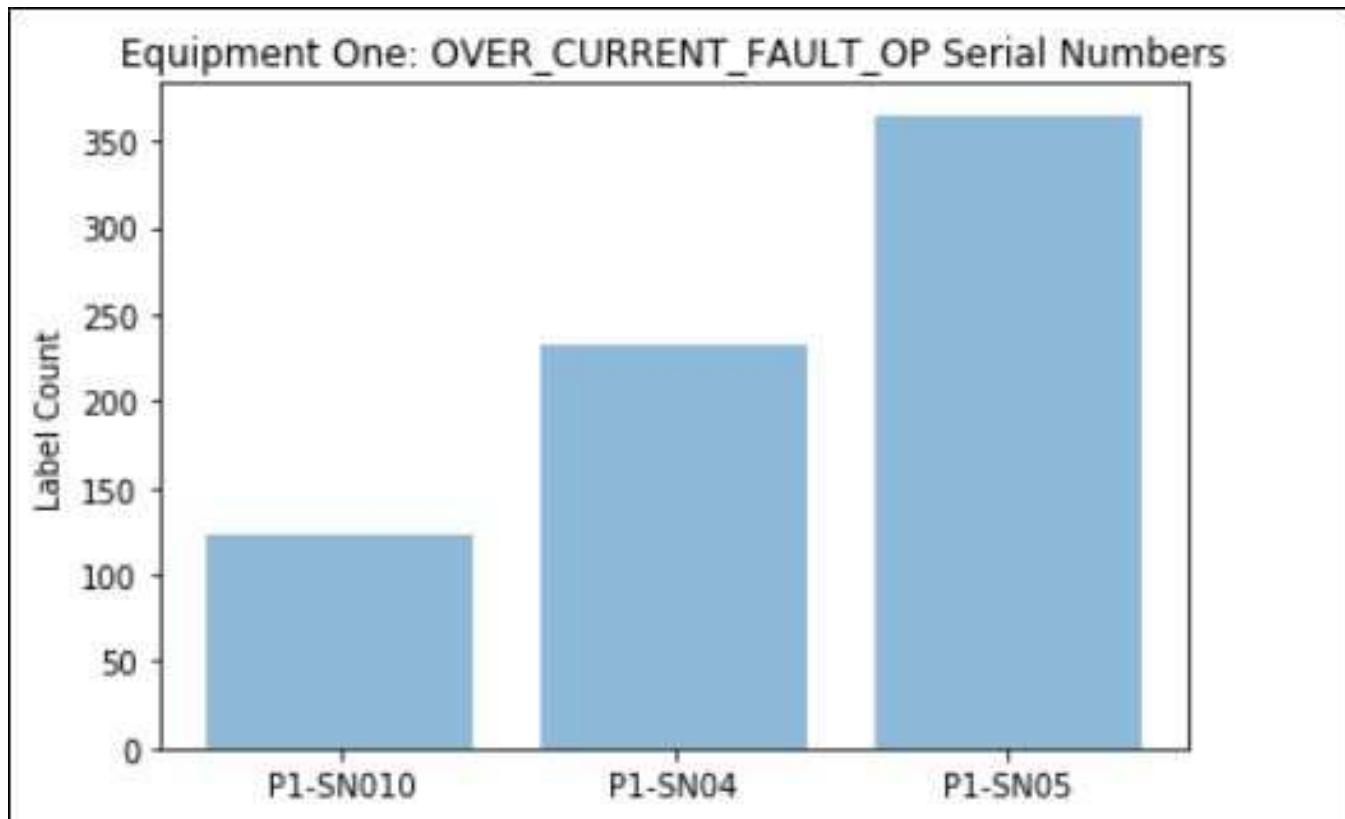


Figure 8: Over current Fault

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-12-16 09.38.46] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.0s.
```

---

```
Compliance Report
```

---

```
name: Ashok Reddy Singam
hid: 337
paper1: Nov 01 17 100%
paper2: Nov 06 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

```
-----  
13  
wc 337 project 13 5451 report.tex  
wc 337 project 13 5434 report.pdf  
wc 337 project 13 366 report.bib  
  
find "  
-----  
  
passed: True  
  
find footnote  
-----  
  
passed: True  
  
find input{format/i523}  
-----  
  
5: \input{format/i523}  
  
passed: True  
  
find input{format/final}  
-----  
  
passed: False  
  
floats  
-----  
  
36: The PHM technology can be put within a broader business context by  
relating it to the Product-Service System (PSS) business model.  
PSS can be defined as an integrated combination of products and  
services where the emphasis is put on the ‘‘sale of use’’ rather  
than the ‘‘sale of product’’ \cite{Tonci2009}. Central to this new  
business model is a shift from selling a product, and its related  
spare parts as required, to selling a solution that supports  
customer needs in the form of a service delivering a fully  
maintained and useable product \cite{Tonci2009}. As shown in  
Figure \ref{F:Figure1}, There are several wireless technologies  
such as 802.11, cellular, and short distance wireless protocols  
can be used to collect and send data to the centralized  
servers. Also, data can be stored in cloud based technologies such  
as AWS, Microsoft Azure, IBM and Google etc. for processing.
```

```

39: \begin{figure}
40: \includegraphics[width=1.0\columnwidth]{images/system_architecture}
}
41: \caption{System Architecture} \label{F:Figure1}
68: The PHM application design process is shown in Figure
\ref{F:Figure2}, which describes various steps of the processes
involved. For the implementation of this project, the sensor
generated data is simulated using SQL scripts due to development
time constraints. However, a detailed step-by-step approach is
provided if we need to plug-in the sensor modules in to the
application.
72: \begin{figure}
73: \includegraphics[width=1.0\columnwidth]{images/phm_process_1}
75: \label{F:Figure2}
129: \begin{figure}
130: \includegraphics[width=1.0\columnwidth]{images/ttf_1}
132: \label{F:Figure3}
135: As shown in the Figure \ref{F:Figure3}, the time to failure
prediction will be estimated based on the classified result data
set and alert the stakeholders to take relevant actions. The
target alert zone will be identified based on the abnormal
behavior of the equipment over the period.
199: Based on the performance as shown in Figure \ref{F:Figure4}, kNN
has been selected to use for this application.
201: \begin{figure}
202: \includegraphics[width=1.0\columnwidth]{images/algperformance}
203: \caption{Algorithm performance} \label{F:Figure4}
249: In this application, an average k-fold cross validation accuracy
of 0.99 percent achieved, which is explained in the appendix
section of the report. Figure \ref{F:Figure8} shows
Classification and Confusion report output obtained from the KNN
model we used for this project.
251: \begin{figure}
252: \includegraphics[width=1.0\columnwidth]{images/knnclassification}
253: \caption{KNN Classification and Confusion matrix report}
\label{F:Figure8}
269: There can be more application specific information obtained from
classified data set to take various decisions. Figure
\ref{F:Figure5}, Figure \ref{F:Figure6} and Figure
\ref{F:Figure7} show various PHM data analytics for this project.
Figure \ref{F:Figure5} displays all the serial numbers of
equipment 1 with bearing degradation problems. The X axis gives
serial numbers while the Y axis gives number of occurrences of
bearing degradation for that particular serial number. Similarly
Figure \ref{F:Figure6} and Figure \ref{F:Figure7} give the
details of over temperature and over current faults of various

```

```
    serial numbers.  
276: \item Bearing Degraded Serial numbers for Equipment Part Number1:  
    Figure \ref{F:Figure5}  
277: \item Over Temperature Fault Serial numbers for Equipment Part  
    Number1: Figure \ref{F:Figure6}  
278: \item Over Current Fault Serial numbers for Equipment Part  
    Number1: Figure \ref{F:Figure7}  
281: \begin{figure}  
282: \includegraphics[width=1.0\columnwidth]{images/DEGRADE}  
283: \caption{Bearing Degradation} \label{F:Figure5}  
286: \begin{figure}  
287: \includegraphics[width=1.0\columnwidth]{images/OVERTEMP}  
288: \caption{Over Temperature Fault} \label{F:Figure6}  
291: \begin{figure}  
292: \includegraphics[width=1.0\columnwidth]{images/OVERCURR}  
293: \caption{Over current Fault} \label{F:Figure7}
```

```
figures 8  
tables 0  
includegraphics 8  
labels 8  
refs 9  
floats 8
```

```
False : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check  
passed: True
```

```
When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====

```
The following tests are optional
```

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# **Big Data Application in Precision Medicine and Pharmacogenomics**

Budhaditya Roy

Indiana University

School of Information and Computing

Bloomington, IN 47040

royb@indiana.edu

## **ABSTRACT**

This article focuses on the impending impact of big data analytics improving health, preventing and detecting illness at a preliminary stage of illness and personalize interferences. The complexity and diversity of biological data are pouring the need of big data analytics and how it is applied in biological field especially in Pharmacogenetics, personalized/precision medicine. Big data is particularly very useful in the healthcare industry as a whole for its data handling intensive nature. Over the past decade, electronic health records (EHR) have become an extensively accepted in hospitals and clinics worldwide and the amount of information is generally daily from a single patient is increasing day by day. Important clinical acquaintance and a deeper understanding of patient disease patterns can be deliberate from such data. It will help to improve patient care as well improve efficiency of patient care and disease prevention. There are few applications pointed out to be effective using big data such as Healthcare data solutions and big data in cancer therapy, continuous monitoring of patients symptoms, healthcare intelligence, fraud prevention and detection. Many people heard about the proposition of precision medicine in State of Union speech of President Obama in 2012. Since then the revolutionary process of precision medicine started to grow rapidly in healthcare industry. On January 30, on the same Precision based medical initiatives, the Obama administration exposed facts about the Precision Medicine tentative plans. Threw with a 215 million dollar investment in the US President's 2016 budget,, the Precision Medicine Initiative will product a new model of patient driven research that eventually support delivering the right treatment to the right patient at the right time[9]. On March 11, 2015, it is reported that China is planning to invest 60 billion Yuan almost 10 billion) in precision medicine (20 billion from the Central Government and the remaining 40 billion from local governments and companies) before 2030 [[9]]. There is a similar necessity of big data application to this latest emergence of biomedical domain. . Big data in precision medicine is the most widely used methods in precision and personalized medicine which is a life changing event in healthcare industry. Personalized medicine or called as Precision medicine is product and services that leverage the science of genomics and proteomics and take advantage of on the trends concerning wellness to enable preventive care. By using big data analytics, prevention and detection of diseases are in a new era of healthcare which essentially improving daily life of every patient. Personalize medicine is the in an era of new modern healthcare innovation. The role that big data analytics may have in interrogating the patient electronic health record headed for improved clinical decision support is discussed. In this paper we try to examine developments in pharmacogenetics

that have enflamed our appreciation of the reasons why patients respond inversely to chemotherapy in cancer treatment. We also try to measure the development of online health infrastructures and the way healthcare data may be capitalized in order to detect public health warnings and control or comprise epidemics. Finally, this paper talks about how a new generation of body sensors in form of implanted in human body may improve comfort, rationalize management of chronic diseases and progress the superiority of surgical implants which could be effectively used in near future. So, let's talk about what is precision medicine? How is it related to other dealings such as personalized medicine and omics technologies (especially in Pharmacogenomics and in Pharmacoproteomics)

## **KEYWORDS**

Keywords- HID348, Precision Medicine, Pharmacogenomics, Pharmacoproteomics. Big data Application, Data analytics, Big data infrastructure

## **1 INTRODUCTION**

The complexity, diversity, and rich context of data being generated in healthcare are driving the development of big data for health [3].The data captured at these portals can also help significantly reduce the cost of drug discovery as improved predictive analytics to determine which drugs work well and which are not as effective for certain conditions. Big data analytics may even allow for uploading the genomics of large populations that can be warehoused for researching new generations of drug remedies. Big data analytics is becoming increasingly popular in modern world with almost every domain. Big Data means lots of data used for analysis and get the insight from the data. Big data applications in general is applicable to any domain such as Retail, Healthcare, Finance and Supply Chain efficiently but in current world the application of Big Data has a major impact on healthcare sector where daily volume of data quadruple in every minute around the globe. Organizations are using Big Data to envisage the future with the goal of making them smarter and competitive in daily work. Applications from Big Data has become from retail industry where Big Data helps retailers gain insights into the customer needs and by monitoring customers' habits future can be effectively utilized, HealthCare and Hospitality[3]. Government agencies are progressively integrating Big Data analytics to control crime and sustain law by foresee the circumstances and by using social media, it is trying to achieve other benefits out of it. So, to get actionable data and perform analytics requires specialized tools which can handle this

massive amount of data as well as help in analysis of the information. There are thousands of Big Data tools available in the market right now which contribute significantly to healthcare analytics. There are open source tools like Hadoop, which is named as big data umbrella and in big data ecosystem. Today's healthcare data are beginning analyzed using aforesaid big data tools such as Pig, Cassandra, MongoDB and others. The quality of health care services in US and across the globe have been enhanced tremendously because of the advancement in health care services, advancements of technologies and Artificial intelligence process which improved the accuracy of healthcare as a service to next level. According to Google Trends analysis, the number of searches using the keyword "fibig data" started to increase dramatically in 2011 and reached the peak in 2017 [3]. Although the term "fibig data" resonances as if it is connected to the area of data science, big data and data science both play a significant role in healthcare research especially in precision and emergency medicine. Conventionally, scientists have adopted the traditional 4Vs criteria to describe big data: volume of data, velocity of data which is speed of incoming and outgoing data and variety which is range of data types [5]. However, from the perspective of medical science, this classification may not be real world sufficient as the 3Vs criteria are forceful and time-reliant. Big Data is a capacious collection of data that cannot be achieved by traditional database management systems (RDBMS). Big Data is an umbrella term used for the enormous amount of data produced from countless of sources such as mobile, web, sensor devices, enterprise applications and rigorous digital repositories. In big data umbrella, data can be structured as well as unstructured or semi-structured. The data varieties from terabytes to exabytes of data [5]. The relational database management systems (RDBMS) have proven inefficient to handle such huge volumes of data in form of patient records such as X-ray, Scan and routine checkup results. Another important factor which renders the conventional database systems inappropriate is that the majority of data being generated as unstructured, the RDBMS systems are only adept to handle structured relational data. Hereafter new tools and systems for data analysis and management have emerged. Volume, velocity, variety, veracity, variability, and value are the three must have Vs of big data and these are condensed in the integral challenges of biomedical and health informatics. Effective ways of confronting these challenges would cover the way for more intellectual healthcare systems focused on early detection, prevention and personalized treatments. As Big data is characterized by the 4 Vs. We discussed about the 4Vs below which essentially contributed to the success of healthcare management [3]. 1. Volume- As data are increasing day by day, it is always in light of voluminous collection of data. The complete volume of data generated these days by real-time applications such as X-ray machines and MRI systems and other external data sources such as Facebook comments, tweets or even patients' data, it runs to petabytes and exabytes of data. Big Data technology empowers us to store this amount of data on dispersed systems [[5]]. 2. Velocity- is the proportion at which data is arrived. As an example, in whole gene sequencing process, one sequence generates huge volume of data and when the sequencing process is completed data arrives at a very higher speed having few time to store and analyze the data. 3. Veracity- When the volume increases so does the quality. Veracity refers to the quality of the data.

There are doubts of good quality of accurate data being generated in recent times. Big Data applications empower working with data which are large in volume, accurate and insightful [5]. 4. Value- Everything in the world has a value so does the data. There is an intrinsic value that the data holds and discovers for analysis. Value is the heart of Big data analytics and the way data generated value in healthcare is just enormous. Modern technologies have made it possible to find the insight from data. Since data is huge and storage capacity leads to an expensive turnaround. Apache Hadoop is the savior in these kind of applications by processing gigabytes of data in a very short span of time and Hadoop ecosystem consisting of MapReduce a different language processing system or Hive and Drill an analytical SQL platform on Hadoop or Spark, in memory data flow system or HBase/MongoDB in memory database systems or HDFS, capable of storing petabytes of data or streaming systems such as Apache Storm and Kafka, overall these all highly capable tools can be profoundly effective in healthcare biomedical data analytics.

## 2 WHAT IS PRECISION MEDICINE

Precision medicine in broader terms is another name of preventive medicine. According to the Precision Medicine Initiative and American Healthcare Association, precision medicine is an emerging approach for disease treatment and prevention that takes into interpretation of individual heterogeneity in genes, environment and routine and lifestyle for each person. This approach will allow medical professionals and researchers to predict more accurately about the treatment and prevention approaches for any particular disease about its effectiveness. It is in divergence to a one-size approach in which disease treatment and prevention strategies are developed for the average person with less deliberation for the genetics-based differences between each individual. Though the term precision medicine is relatively new in medical industry, the concept has been there and as a part of healthcare for many years. As an example, a patient who needs a blood transfusion is not given random blood from blood bank storage rather it will be from a process of donor's blood type is matched to the recipient to decrease the risk of future problems [10]. Although illustrations can be found in numerous areas of medicine the role of precision medicine in day-to-day healthcare is relatively restricted. Medical researchers expect that this approach will increase in many areas of health and other healthcare domains in upcoming years. Precision medicine is being sought to transform how we as a whole improve health, treat and prevent disease. Today most of the medical treatments are intended for the average patient using the one-and-one approach. However, in many cases, this approach is not at all effective because treatments can be very successful for some patients but not for every patient [5]. As an example, if Patient A and Patient B both have stage 3 lung cancer, giving the same chemotherapy to both the patient helps one but not the other. In precision medicine with the help of big data technology, medicines are targeted to specific genomic sequences rather than a random selection. In advanced countries like USA, a rigorous process is already in place to target particular genes after finding the root cause of the disease. It is a big data application which enables to store the data and use analytical tools to get useful information out of it. Overall, Precision medicine is a field

of medicine that takes into interpretation individual differences in people's genes, environments, microbiomes, habitual effects, and family history to make diagnostic and beneficial strategies accurately personalized to individual patients. Precision medicine is a newer term referring to a similar ground compared to another word if personalized medicine. The term if precision medicine arrived the scientific dictionary in the year 2008 when business strategist Clayton Christensen, of Harvard Business School in Boston, invented the appearance to describe how molecular diagnostics allows physicians to unambiguously diagnose the cause of a disease without having to rely on perception [5]. The name precision medicine didn't gain enough attention until 2011 when a committee convened by the US National Research Council placed out a plan for modernizing the classification of disease on the foundation of molecular information such as causal genetic variants instead of a symptom based cataloguing system. The committee called the report Toward Precision Medicine [3]. There are many areas where precision medicine is vastly applicable and are very much beneficial such as, finding correct dose of prescription drugs, root cause analysis of a disease and so on. The field of pharmacogenomics aims to understand how genetic variations influence individual responses to medications. Genetic tests for supervisory treatment decisions are becoming increasingly available across miscellaneous areas of medical care. These kind of tests provide more effective drugs to patients earlier in their treatment and with fewer negative side effects and in less costly than previous tests. Precision medicine is also pertinent in Cancer detection, Genomics and cure process. Oncology is the target of some of the most auspicious precision medicine approaches available today. Cancer forms through the gradual accumulation of genetic DNA changes in genes that regulate cell growth. That is why, cancer is very much an illness of the genome. Depending on where in the body the cancer started and the types of genetic changes the cells grow, different types of cancer have very different genetic profiles which completely varies person to person and highly dependent on their family history. These genetic sequences can be used in a number of ways to help medical professionals choosing the best treatments for each individual patient. Growing tissues replacement is another way to apply precision medicine in pharmacogenomics [3].

## 2.1 Personalized Medicine

The concept of personalized medicine dates back many hundreds of years although the term seemingly similar meaning with precision medicine. Mere from 19th century, scientists started to measure the chemistry of root cause of any illness and the research improvements are granular over time. With the growth of the pharmaceutical industry and medical technology industries in recent times came the rise of genetics, data mining and imaging. Halfway over the period, comments of specific alterations in retort to drugs contributed growth to a body of study attentive on classifying crucial enzymes that play an important role in disparity in drug absorption and reaction and this is helped as the basis for pharmacogenetics. In recent times, sequencing of the human genome customary in motion the transformation of personalized medicine from an knowledge to a practice. Personalized diagnosed tools are now created with rapid developments in genomics along with advances high critical areas

such as computational biology, medical imaging, and regenerative medicine and treatment [5]. Personalized medicine first appeared in available mechanism in 1999 with the creation of some of the domain specific core concepts even dating back to 19th century [2]. So basically personalized medicine is referred to treatment depending on each individual's personal structure and history. Initially, personalized medicine is the idea that assortment of a treatment should be custom-made giving to the individual patient's specific physiognomies, including age, sex, gender, height, ethnicity, diet, and environmental factors against traditional clinical trials on group of people which has been happening since the invention of medicine happened [5]. Scientists got interested on personalized medicine when medical professional started understanding the essence of gene in human development. Several human genome projects have been conducted since then and the importance of personalized medicine started in limelight. With deceitful out in order the 3.2 billion units of our DNA, scientists flashed a blaze of detection and a detonation of genomic knowledge in medical science history [2]. Novel omics technologies including microarrays, whole genome single nucleotide polymorphism [SNP] chips, RNA interference high-throughput transmission, next generation sequencing are the few procedure which accompanied with this revolution. All the above launch a new epoch in personalized medicine which is called genomic revolution era which bids us limitless probable and countless promise containing the expansion of personalized medical products for each individual based on their sole genomic information [10]. Advancement of genomics science along with the developing of new omics technologies, personalized medicine is today frequently well-defined as a combination of molecular profiling (omics methods) and customary methods such as family history, lifestyle and environment, which create analytic and beneficial strategies precisely personalized to individual patients [2, 5].

## 3 BIG DATA IN PRECISION MEDICINE

Once again the term Big data is signifies in collection of large and complex data sets which are difficult or sometimes impossible to process using common database management tools or traditional data processing applications even with modern advancement of traditional data warehousing tools such as Amazon Redshift. In 2012, the Obama administration announced the Big Data Research and Development Initiative [7], which explored how big data could be utilized to address important problems faced by the overall healthcare system. Since then, Big Data becomes such a big term that people tend to claim any kind of data analysis to be if Big Data if characterization. The overall concept of big data can be explained in various ways. One way is, Big data is a comprehensive term for any collection of data sets are so voluminous that processing the data in the begging stage itself is very hard. With four if Vif characterization of big data m, complexity arises more to collect data and make meaningful information out of it. Omics data, mobile internet real-time data and electronic health record data are the top three areas for Big Data in medical research. Precision medicine will use all of these three Big Data. In fact, among the 215 million investment in the USA President's 2016 Budget, 130 million (over 60 percent) will be used for building a large US cohort for precision research [7]. In this regiment study, the scientists will use widespread omics

data, electronic health record data gathered from several hospital and private practices along with mobile internet data [\*]. Thus, omics and medical big data are one of the key pairs in the success of precision medicine in healthcare industry as a whole.

## 4 BIG DATA CHALLENGES IN HEALTHCARE

- Whenever anything benefits us, that comes with its own challenges and problems. The primary idea of big data to be applied in healthcare is to roll massive healthcare dataset with individual information. As the need of more data driven enterprise grows Besides general challenges inherent to the analysis of big data such as missing data, vague data, and varied data, employing big data in health care systems imposes new challenges which includes the lack of reliability and a solid data governance of some biomedical data, issues of privacy and security and confidentiality, insufficient data from random clinical trials including successful and failed trials, and overall low quality data. Challenges in machine learning and statistical applications also put the analytics in challenging situation where model development and execution are critical to success[2]. Healthcare providers who have hardly come to grasps with driving data into their electronic health records (EHR) are now being questioned to pull actionable insights out of them and apply those learnings to complex initiatives that straight impact their repayment rates. Organizations who can integrate this data driven technological innovation to their healthcare operations are in the most benefit[6]. Data assets and data insights can be achieved by using healthier patients, increased visibility in operational excellence, lower care costs and higher staff and consumer satisfaction rates are among the many benefits of turning data assets into data insights. The journey to evocative healthcare analytics is difficult challenge and problems by solving those will benefit the industry to the highest extent. The way overall big data analytics work, collecting, storing, analyzing the data require clear presentation to the staff members to understand the overall workflow process[5]. Analyzing genomic data is a computationally are some of the top challenges organizations typically aspect when striking up a big data analytics program and how can organizations overawed these issues to attain their data driven clinical and financial goals are the most important aspect of big data implementation. Understanding unstructured clinical nodes, storing unstructured patients health records are complex in nature and specialized training is required in implementing the analytics platform is essential. Some of the pitfall of big data application in precision medicine is discussed below.

### 4.1 Data Collection

This is the most crucial stage in any data driven technologies, capturing the patient's behavioral data through several sensing processes; with their numerous social interactions and communications. The data many come from many sources or in different format but not everywhere data governance is properly applied while collecting the data. Capturing data which is clean, comprehensive, correct, and well formatted for use in diverse systems is an ongoing combat for organizations, many of which are not on the endearing side of the battle. As an example, electronic health record capturing in right movement help physicians to access the accurate picture

of the patient's history. Oftentimes, delay in collecting this data create problems which eventually leads to unhealthy environment and future risks. Revolving Healthcare Big Data into Actionable Clinical Intelligence Providers can start to recover their data capture procedures by ranking valuable data categories for their specific plans, conscripting the data governance and honesty knowledge of health information management professionals and evolving clinical documentation improvement programs that tutor clinicians about how to confirm that data is valuable for downstream analytics [5]

### 4.2 Data Cleaning

Healthcare providers are well familiar with the importance of cleanliness in the clinic and the operating room but they are not aware on many things which could lead to a clear picture of the meaningful data. Data which is dirty and raw might have a potential impact on big data analytics projects and can screw up the true insight completely. Data cleaning also known as data scrubbing always ensures that data is not inconsistent, proper and useful in perspective and predictive analytics. Though when everything started, data cleaning was a manual process, but now with the help of big data quality tools, cleaning data has been easier than ever before. Since data cleaning is complex and tedious process in particular healthcare system, oftentimes big data analytical tools stand by the first door where data streamline occurs. Which eventually cleans the data with a global standard before it entered to main stream pipeline.

### 4.3 Data Storage

This is the most critical place where big data application play a key role. As the volume of healthcare data grows exponentially many healthcare providers are not able to manage the costs and effects of on premise data centers. Although many organizations are most happy with on premise data storing which also leads to security issues and data governance issues. With the help of cloud storage almost 90 percent of healthcare providers have chosen cloud based data storage centers which provides better flexibility and availability of data. Amazon web services, Microsoft Azure cloud and Salesforce cloud have put the data storage industry to the utmost point where no longer organizations need to worry about the cost and capacity of storing humongous amount of data. The cloud offers sprightly disaster recovery, lower set up and upfront costs and easier development, although organizations must be extremely careful about choosing partners that understand the significance of HIPAA law and other healthcare related compliance and security issues [3]. Many organizations finish up with an amalgam approach to their data storage agendas, which may be the furthestmost supple and workable approach for providers with variable data access and storage necessities. When creating hybrid substructure providers should be cautious to safeguard that dissimilar systems are able to interconnect and portion the data through extra segments of the organization when necessary [6].

### 4.4 Data Security

Data security is the number one priority for healthcare organizations, particularly in the wake of a hundreds of data breaches, hacking, and intrusion incidents. Data is so sensitive especially

in healthcare systems that a proper security measure has to be taken to protect the data. Healthcare privacy law such as HIPAA and others put the organizations in the front door where every healthcare providers must conform the law to protect the data. For precision based medicine era, this has become more important with each individual patient's data being captured and analyzed. Since genomic science is completely depending on data architecture, one data breach can push the healthcare provider in a tremendous reputation and financial loss. Due to this, security is one of the most talked topic in personalized medicine [2].

#### 4.5 Data Governance

Healthcare data, particularly on the clinical side has a long ledger life. In accumulation to existence required to keep patient data available for at least six years of time frame, providers might request to use de identified datasets for scientific projects, which makes continuing stewardship and curation an important concern [2]. Any data can be used for variety of other purposes as long as data masking is properly applied to the dataset. Understanding of the data when it is created by whom and for purpose can lead to positive results while in research.

#### 4.6 Data Querying

Vigorous metadata and robust stewardship procedures make organizations to comply with data querying very effectively. There are many tools in the market which can give access to query from databases to get the useful information. Azure datalakes, different programming based API, Hadoop Sqoop are the few tools which help in big data query language extraction. Many organizations use Structured Query Language (SQL) to dive into large datasets and relational databases, but this can only be true if end user can trust the data they are working on which can provide useful information to them [2]. Data Reporting: Reporting is the end to end product of any data collection and process. Big data reporting can help transform virtually all aspects of the enterprise. From quickly producing actionable intelligence to driving productivity to gain real time visibility into customers and markets, big data analysis and big data reporting promise to deliver a wealth of benefits for competitive advantage [\*]. Many companies including not for profit hospitals use reporting as their sole decision making procedure. Data reporting helps unveils the insight into charts and graphs and visualize the insight to the target audience. In healthcare organizations especially in precision medicines, reporting of the finding is must have to take informed decision. Data reporting has the potential to show about the result of an ongoing research study or data findings. [2].

#### 4.7 Data Visualization

In patient care, a clean and attractive data visualization can make it much easier for a clinical staff to understand the fundamental very easily and take decision based on it. Color visualizations are a popular data visualization technique that typically yields an immediate response as an example, red, black color divergence is. Organizations must also consider good data presentation practices such as charts, graphs, scatterplot. Common examples of data visualizations include heat maps, bar charts, pie charts, scatterplots,

and histograms, all of which have their own specific uses to prove concepts and material.

#### 4.8 Data Update

Healthcare data is non-static and almost all the elements requires an update in daily interval. Some datasets such as patient vital signs and symptoms, may require frequent update. But patient's demographic information may change once in a while. Since in genomic since changes are captured in every interval or procedure, there has to be constant update to the proper and existing dataset. The most critical phrase comes when incremental data is gathered and new data is added to existing dataset. In precision medicine, medical professional compares whole gene sequences in different timeframe of the disease and in different medicine stages. In these case, data has to be updated regularly in order to analyze the proper data [? ]. Organizations should also confirm that they are not making needless identical records when endeavoring an update to a single component which may make it problematic for clinical staff members to access needed information for patient decision making.

#### 4.9 Data Sharing

With the essence of electronic medical record, data sharing become easier but complicated in healthcare analytics. With large volume and the structure of the data, healthcare providers and researchers are immensely beholding data which may contribute finding to their scientific invention. Data exchange is a perpetual worry for organizations at any costs. With the increase data volume and nature of the data, it is getting more and more difficult for the organization to move data from one to place to another without losing information and change in pattern on data lineage can lead to significant mislead information. [6].

### 5 PRECISION MEDICINE AND OMICS

With the growth of big data, organizations move into NOSQL databases where security is a growing concern. Though we found there are severe security issues in most of the NOSQL databases which are used today in big data environment. Lack of security measures put extra sensitivity to the overall big data applications being NOSQL databases are heart of any big data project. Though not reached at pick, constant evaluation and research are in process to make NOSQL databases more secure in near future. The evolution of omics outlining technologies significantly benefited studies are conducted on diseases mechanism, molecular diagnosis and personalized treatment [5]. The study of omics is strongly related to the study of biology as a whole and precision medicine. There is a strong connection between Omics and Precision medicine and big data as a whole has become the core of precision medicine. The advancement of precision/personalized medicine depends heavily on the ability to acquire biological aces at omics interval though the training of precision medicine does not use sole omics data and omics knowledge [1]. This happens due to molecular characteristics found from omics data can categorize diseases and classify population of patients appropriate to assured common treatment more exactly [5]. Biology has become more data intensive and technological intensive subject .Following this trend, many of the emerging fields of large-scale data rich biology are designated by

adding the suffix *fi*-omics to previously used definitions. Particularly, the word omics refers to a field of study in biology ending in the suffix *fi*? omics and it is related addresses the objects of study of such a field[5][2]. Pharmacogenomics is the study of how a person's response to drugs is affected by his genetic makeup [3]. It combines pharmacology which is also called the science of drugs and genomics which is the study of gene and their functions to develop effective, proper medications that will be personalized to a person's genetic makeup. Pharmacoproteomics, essentially a sub discipline of functional pharmacogenomics which is a study of how the protein content of a cell or tissue changes qualitatively and quantitatively in response to treatment or disease, what the protein-protein and protein ligand interactions are in related to drug response, and how a person's protein variants in quality and quantity affect a person's response to a drug [10]. In modern days, the pharmaceutical industry has developed strong interest in Pharmacoproteomics with the anticipation that this technology will lead to the empathy and authentication of protein targets and eventually to the detection and growth of feasible drug candidates. Pharmacogenomics and Pharmacoproteomics will help the prescription of drug and related doses to a patients based on response to a drug which greatly indorsing the advance and practice of precision/personalized medicine [10]

## 6 MANAGEMENT AND PROCESSING OF OMICS DATA

There is no shortage of data in healthcare and it is growing 40 percent annually according to IDC. Data in healthcare is not always about volume in healthcare but there are several factors can contribute to it as well such as *fiRegulations*, *fiComplexity* and *fiIntegrity*. To process the volume and complexity of Omics data, there is a need of major investments in research laboratories in form of computational and storage capabilities. Laboratories need servers or cloud service storage access to store this massive amount of data. In traditional way, servers are costly in maintenance and ended up in profligate or sub optimal servers which are even more added cost load. In the past decade, cloud computing is closing the gap in handling omics data. Cloud computing is a high scalable multi-processing semantic environment which operate virtually with some of the great benefits to any organization such as costs, speed, global scale, productivity, performance and reliability. A good example comprise the *fiEasyGenomics* cloud in Beijing Genomics Institute (BGI) and *fiEmbassy* clouds as part of ELIXIR project in collaboration with multiple European countries (UK, Sweden, Switzerland, Czech Republic, Estonia, Norway, the Netherlands, and Denmark) [8]. In several circumstances, Graphic Processing Units or GPUs are also used in cloud environment as GPU's are named to provide faster processing of data compared to Central Processing Units or CPU's. There is also a high need of parallel computing in processing of Omics data along with data validation need in research platforms before Omics data are utilized. Along with software application in storage which is only one side of the medal there is a huge need of integration between biological system and the Omics data. Moreover, analysis of big data requires most recurrent data access to turn data into real knowledge. Even though we can take up the occurrence of an appropriate volume

of bandwidth inside a cluster, there is a great need to have use distributed computing infrastructure in effective solution adaption. In big data technological umbrella there are two highest performance parallel file systems are the General Parallel File System (GPFS) [8] a product of IBM, and Lustre [] which is an open source platform. Predominantly most supercomputing systems such as Lustre is used in Titan, the second supercomputer of the TOP 100 list (December 2017) [8]. The storage capability of Titan contains more than 20,000 disks which is equivalent to 40 Petabyte of storage and almost 1 Terabyte per second of storage bandwidth. Among many software companies in big data business. IBM file management solution is operational as a regulator plane for smooth data handling. Since the access of data is so frequent, modern software's can automatically switch less frequently accessed data to the less expensive storage available in the infrastructure keeping the most expensive storage for critical and sensitive data. Nowadays, moving the data from less expensive storage to expensive storage is completely depending on analytics supported decision making which includes pattern, storage characteristics and network pattern. Hadoop file systems plays a very important role in Omics data storage and overall data processing capabilities as explained above. Besides HDFS, Middleware is also essential in development of user specified custom solutions. A suitable example is R, considering a statistical programming tool, R is more robust now to handle high volume of data in biomedical data analytics and with help of middleware it can be used best in Omics data analysis as well.

## 7 ANALYSIS OF BIG DATA IN OMICS USING COMPUTATIONAL FACILITIES

HPC clusters along with grid computing used as customary platform for big data applications. Over the years it has been documented so many drawbacks of these platform in real environment where data manipulation and data integration were critical for success. Through cloud computing small and medium size laboratories can now leverage the power of data by accessing the data they want through cloud storage devices. Cluster computing which is a data parallel approach processes data independently with a high scalability. De Nova assembly algorithm is a renowned algorithm which is developed on cluster computing. This system can process daily genome sequence and find sequenced reads overlaps using in memory distributed approaches. Here the data is held in memory clusters unless it is sequenced differently [4]. Another high computing big data application is Intel's Xeon Phi which can deliver massive parallelism and vectorization to support high performance computing applications. Xeon Phi is an excellent support systems application in data discovery workloads and high dimensional matrices can be used to a level which can significantly benefit the thread level parallelism. Along with some established and newly added big data application which can great change the biomedical industry, we have documented the used of data semantics in many research laboratories in their data discovery process. Semantic data uses RDF or Resource description framework, Web Ontology Language or OWL, SPARQL or Protocol and query language for semantic web data sources and extensively use of XML (Extensible

Markup Language). Oftentimes these have a dense use in bioinformatics and biostatistics to integrate all data formats and standardize existing ontologies [4].

## 8 CONCLUSION

Personal medicine, Omics technologies and pharmacogenomics are the evolutionary invention in medical industry, holding the hands of these medical concepts scientist can not only find the cancer cell in human body parts as well as start of cancer as a disease in a particular cell. These all is possible due to the essence of big data which not only helped organizations to tackle the voluminous data effectively but to use them in a way to get meaningful insight out of it. Massive parallel computing and clustering are now opened up new window in medical research where processing of huge amount data is better than ever before as well as build automated model on top of it. Whole gene sequencing is an example of how a big data can help strong millions of genetic information in a single storage system and take useful information out of it. With the help of big data in precision based medicines scientists are now able to predict the origin of the disease, track and cure it more effectively.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and I523.

## REFERENCES

- [1] Shein-Chung Chow1 and Fuyu Song2. 2016. *Some Thoughts on Precision Medicine*. Journal of Biometrics and Biostatistics, NA, Chapter 1, 1.
- [2] Mohit Dayalfk and Nanhay Singh. 2012. *Indian Health Care Analysis using Big Data Programming Tool*. NA, Web, Chapter 1, 2. NA
- [3] Andy Futrel. 2012. *Building Genomic medicine capability* (1st. ed.). 4th, Vol. 2. MD Anderson Cancer Research center, Boston, MA, Chapter 1, 1. [https://doi.org/10.1007/978-1-4614-0923-7\\_4](https://doi.org/10.1007/978-1-4614-0923-7_4)
- [4] Sandra Gesing and Daniels DifAgostino Ivan Merelli, Horacio Prez-Sánchez. 2014. *Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives*. 1st, Vol. 1. BioMed Research International, Web, Chapter 1. <https://doi.org/NA>
- [5] Daniel Richard Leff and Guang-Zhong Yang\*. 2015. . 1, Vol. 1. Engineering.org, NA, Chapter 1, 2. <https://engineering.org>
- [6] IEEE Chih-Wen Cheng Member IEEE Chanchala D. Kaddi Member IEEE Janani Venugopalan Member IEEE Ryan Hoffman Member IEEE Po-Yen Wu, Member and IEEE May D. Wangfk, Senior Member. 2016. *Omic and Electronic Health Record Big Data Analytics for Precision Medicine*. NA, Web, Chapter 1, 1. <https://doi.org/10.7/3-105.876>
- [7] White House Press Release. 2015. *Precision Medicine Initiatives*. White House Press Conferences, NY, Chapter 1, 1. <https://doi.org/10.1007/978-1-4614-0825-3>
- [8] Marx V. 2013. *The big challenges of big data*. Nature. BMC Medical Genomics, Web, Chapter 1, 1. [https://doi.org/498\(7453\):255ff?160](https://doi.org/498(7453):255ff?160)
- [9] Whitehouse (Ed.). 2012. *Precision medicine Initiatives*. 1st, Vol. 1. Purdue University Press, Purdue University, Indiana, Chapter 1. <https://doi.org/NA>
- [10] Xiaohua Douglas Zhang. 2015. *Pharmacogenomics & Pharmacoproteomics*. 1, Vol. 1. Merck Research Laboratories, Web.

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
I was expecting a ',' or a '}'---line 62 of file report.bib
:
: chapter = "1",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
I was expecting a ',' or a '}'---line 125 of file report.bib
:
: doi = "10.7/3-105.876",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
Unbalanced braces---line 142 of file report.bib
:
: }
(Error may have been on previous line)
I'm skipping whatever remains of this entry
Warning--I'm ignoring editor11's extra "address" field
--line 168 of file report.bib
Warning--I didn't find a database entry for "editor0"
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3085 of file ACM-Reference-Format.bst
```









```
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--unrecognized DOI value [498(7453):25560]
Warning--page numbers missing in editor01
Warning--unrecognized DOI value [NA]
Warning--empty chapter and pages in editor04
(There were 115 error messages)
make[2]: *** [bibtex] Error 2
```

latex report

```
=====
[2017-12-16 09.39.33] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.5 L84 : [editor0] undefined
Missing character: ""
```





```
Missing character: ""
There were undefined citations.
Typesetting of "report.tex" completed in 1.0s.
./README.yml
7:14      warning  truthy value is not quoted  (truthy)
```

---

## Compliance Report

---

```
name: Budhaditya Roy
hid: 348
paper1: 100% Oct 25 17
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
7
wc 348 project 7 579 content.tex
wc 348 project 7 6353 report.pdf
wc 348 project 7 553 report.bib
```

```
find "
```

---

```
103: Do not use "these quotes" but use these ‘‘these quotes’’.
```

```
passed: False
```

```
find footnote
```

---

```
113: \footnote{do not use footnotes}.
```

```
passed: False
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
62: In Figure \ref{f:fly} we show a fly. Please note that because we  
use
```

```
69: \begin{figure}[!ht]
```

```
70: % \centering\includegraphics[width=\columnwidth]{images/fly.pdf}
```

```
71: \caption{Example caption}\label{f:fly}
```

```
86: or generate them by hand while using the provided template in  
Table\ref{t:mytable}. Not ethat
```

```
89: \begin{table}[htb]
```

```
92: \label{t:mytable}
```

```
figures 1
```

```
tables 1
```

```
includegraphics 1
```

```
labels 2
```

```
refs 2
```

```
floats 2
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
106: Do not use Figure 1 user the ref for the figure while using its
      label
passed: False -> labels or refs used wrong
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not cahnge the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
WARNING: figure and below may be used improperly
```

```
67: figure below.
```

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
I was expecting a ',' or a '}'---line 62 of file report.bib
:
:    chapter = "1",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
I was expecting a ',' or a '}'---line 125 of file report.bib
:
:    doi = "10.7/3-105.876",
(Error may have been on previous line)
I'm skipping whatever remains of this entry
Unbalanced braces---line 142 of file report.bib
```

```
:  
: }  
(Error may have been on previous line)  
I'm skipping whatever remains of this entry  
Warning--I'm ignoring editor11's extra "address" field  
--line 168 of file report.bib  
Warning--I didn't find a database entry for "editor0"  
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha  
while executing---line 3085 of file ACM-Reference-Format.bst
```





```
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ivan Merelli, Horacio Prez-Snchez,Sandra Gesing, and Daniel
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ivan Merelli, Horacio Prez-Snchez,Sandra Gesing, and Daniel
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ivan Merelli, Horacio Prez-Snchez,Sandra Gesing, and Daniel
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--page numbers missing in editor11
Warning--unrecognized DOI value [NA]
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Chanchala D. Kaddi, Me
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Ch
```



```
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Po-Yen Wu, Member, IEEE, Chih-Wen Cheng, Member, IEEE, Cha
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--unrecognized DOI value [498(7453):25560]
Warning--page numbers missing in editor01
Warning--unrecognized DOI value [NA]
Warning--empty chapter and pages in editor04
(There were 115 error messages)
```

bibtex\_empty\_fields

---

```
entries in general should not be empty in bibtex
```

find ""

---

```
4: author =      "",  
22: editor =     "",  
40: editor =     "",  
55: editor =     "",  
65: address =   "",  
71: editor =     "",  
85: editor =     "",  
156: url =      "",  
161: editor =     "",  
passed: False
```

ascii

---

```
=====
```

The following tests are optional

```
=====
```

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Diversification of Big Data

Shiqi Shen

Indiana University Bloomington  
1575 S Ira St  
Bloomington, Indiana 47401  
shiqshen@indiana.edu

## ABSTRACT

There are some ideas around the conception of big data and how it is used in the market outside of a programmer's computer. What I mean to try, and state here is that there is an idea of how big data is used and functions in the world. As a programmer, we cannot just focus on what we contribute to the programs, meaning we cannot only look at our computer data and creation of such programs as a means for our own scientific endeavors and egos, we must understand how such creation are reflected into the world and how such programs create a very interesting market exposure. Therefore, we write this paper to analyze the enterprise of big data. We will use this paper to seek out the multiplicity of avenues in which big data is used by our technological world (mainly those that seek to use big data to create diversified consumer experiences).

## KEYWORDS

i423, hid109, Big Data; Social Media; Online, Shopping, Customers; Pricing; Dynamic, Internet, application

## 1 INTRODUCTION

In order to begin a type of observation around this topic, we need to first establish what types of enterprise we will look into in order to make our observations. We will look at Online Shopping, streaming services, and Social media. These forms of enterprise concentrate their use of big data to create a diversified user experience, one which looks to creating more possibilities and other forms of consumer products for consumers. We would also like to use such an observation to understand the use of big data in these types of enterprises and how such uses render experiences and technology of big data as a good form of consumer study.

## 2 OBSERVATION

Before we begin our in-depth observation, it is important to showcase the Big Data technologies, some components of big data that make it what it is. There is of course an array of technologies that facilitate and create big data. In the creations of big data, we can see many enterprises like the generations of web pages (in which individuals, corporations, government, and the like, produce these pages with data). We can also see digital imagers that facilitate data collection as well. These types of data can come from telescopes, MRI Machines, and Video Cameras. Another source of data production can come from biological and chemical sensors, things like microarrays and environmental monitors [21].

From production of data there must be a medium that collects this data, that is of course usually computers. These data can be collected in things like the internet and localized sensor networks.

Qiaoyi Liu

Indiana University of Bloomington  
3209 E 10th St  
Bloomington, Indiana 47408  
ql30@umail.iu.edu

These sources of course can both collect and analyze the data. From collection, there are also storage capacity for these data. In that, we find that there are many storage type disk (magnetic disk) that can hold tons of data. There are also cloud services with which data can be stored.

Big data is crucial to the developments of many businesses that drive on the interactions and recommenders of their clientele. When we are looking to understand big data and how it is used by big contenders, we must first look to the companies to see what uses they have for big data. Let us begin with looking at Amazon, an empire built on e-commerce, the selling and recommending of products to consumers at efficient and effective rates. When we look at how this company succeeds, we can see that the main component to its success comes from its unique and bold use of big data.

## 3 ALGORITHMS, PREDICTIONS, AND BIG DATA CRAZE WITH ONLINE SHOPPING

When a consumer is shopping on Amazon, they can click on the image of the product they would like to view. During this process, they will probably go back and look at other products to compare the product to other products that are similar. Therefore, what can a company do to ensure that they are helping their consumers with this process? By pooling this data into other similar consumer searches and sells. For Amazon to do this they have a very complex system of enterprise that is meant to survey this, their use of big data. But it also comes down to very simple standards that are used by consumers.

When a consumer is shopping on Amazon they can administer filters to help them narrow their searches of products that they like. But during this process of administering these filters, the consumer is also being surveyed of searches made in the past to better associate what the consumers wants and needs within the products that they are shopping for.

This type of use of big data allows for any company (mainly Amazon like companies) to focus their resources on the calculation of products that an individual consumer would want based on searches that they have made. But the best part about this is that the information is pooled on multiple similar searches and consumers to best devise the necessary searches to show a consumer what to buy [25].

But beyond this simple observation of the big data usage, we need to understand what forms of implementation must be taken into consideration when creating these types of programs. Now,

what really needs to be focused on here is the interpretation of the data that is collected by Amazon to provide consumers with their recommendations.

Within big data itself, the name gives away what it is, it is a cluster of data, a large, almost seemingly insurmountable amount. There must be some sort of way to interpret the data results as they come in. For this process, it needs to be understood that this data interpretation cannot happen within a void, rather what is done for these data to be interpreted and re-designated as recommendations back to a user would be through a process of examining detailed assumptions and rethinking the analysis [1].

This process of observing and interpreting big data itself can have issues, such as those of bug interference based on programs that are being used in order to interpret these large data pools, and data can become erroneous. However, a way in which these types of issues can be resolved comes in the form of predetermined data assumptions. Data assumptions that are made to help companies who use big data to narrow in on data pools to create a seamless connection of data gathered to products showcased based on data that is collected, interpreted, and re-designated to users. This is done to help devise the necessary sales goals or recommendation services that come from these companies use of big data.

Also, when we are looking at the process by which data is collected, there will be data that is of no importance to the necessary processes by which the data is collected for. The difficult task becomes that of filtering out the useless information without removing the information that is of importance. To do this, there have been advancements made within the scientific community to reduce the plausibility of such turn out to happen. The process seeks to monitor faultiness that can be caused by sensors to lower the chances of data that is useful to not be discarded alongside data of no importance. This is also where algorithms that are made to establish key components of data come into play as well.

These forms of interpretation of course come from algorithms that are used to detect the data in forms of patterns. A perfect example of this would be comparing both Amazon and Netflix's use of algorithms that recommend to users what to buy or watch next. Machine learning within the recommendation system is what enterprises this use of algorithms. The machine learning itself will compare the histories of purchase and views to establish a statistical model of a collective pool of millions of other users to generate the necessary continued recommendations to users. The use of algorithms and machine learning also helps to establish a base line of what it is that users like and are more attuned. The use of algorithms is to ensure that the data that is being collected is correctly sorted and re-designated to users in the forms of recommendations. This is because large quantities of data require these types of assumptions to calculate and extract knowledge from the data that is collected [21].

As a programmer, machine learning and the creation of these algorithms truly fascinate when they are being used in order to create a recommendation system. Since the collection of data is

pooled, the use of algorithms to reach a particular end, based on data analysis, makes for a very interesting usage. As the use of data algorithms not only seek to facilitate the necessary recommendations, but also filter through millions of data informatics to agitate the necessary products to ensure the transparency of recommendations created. This can also be seen in what was stated above regarding consumers being able to filter products with precise search options. These filters can act as filters for filters. As in the filters chosen by consumers help to facilitate precise sifters and allows for algorithms to pinpoint precise outlines of data that help to recommend even finer tuned recommendations for consumers.

Big data, from this perspective, can be seen as a general tool that can be created to make precise measurements that are used in order to facilitate the necessary recommendations for consumers. This is of course a process that is keyed out by the use of interpretations and algorithms that are necessary to the specifications made on big data that is collected. Therefore, big data is used precisely well as a tool to pool and narrow pattern like data to ensure that completed data of particular patterns can always correctly correlate to consumers as well as viewers who use services such as Amazon and Netflix (This paper will not observe Netflix in full as it did with Amazon since the use of big data analyzes and processes to creating recommendations for consumers are very similar).

Therefore, when we are looking at companies like Amazon and Netflix, we are looking at companies that use big data to create predictive analytics. Predictive analytics has many uses however and cannot just be subjected to Amazon and Netflix's use on commercial needs. The use is mainly attributed to uncovering patterns and highlighting relationships with the data that is being observed. Because of this very nature, big data that goes under predictive analytics are being done soon to search out these two main components. There is also the process of trying to find past data patterns outcome variables and trying to deduce them for the use of the future (observing patterns from a specific time-period to see if such trend continues again at another given future based on certain functions that existed in the data of the past and then comparing that to the future) [14].

Further delving into predictive analytics, which is the main use of big data for our commercial subjects, we see that there are also forms of linear regressions. The use of this is to find interdependencies within outcome variables and explanatory variables in order to use them in the process of making predictions or to right out make the prediction itself. This looks to focus the data that is collected into predictive measures that are precise to the sets of data that are collected and distributed through this type of analysis. Above, when looking at machine learning, this is categorized as a neural network. Neural networks are a collective entity, something like that of the human brain, if an artificial neural network can be defined as a computing system made up of number of simple highly interconnected processing elements which processes information by their dynamic state response to external inputs. Within the neural networks, machine learning works within the sphere of the networks to generate and learn from data collected to predict,

showcase patterns, and classifying input data [22].

To finish up the observation here on big data uses for predictive analytics and the basis of enterprise that is that of Amazon and Netflix's ability to use big data to create such systems, the paper needs to understand a little further into the principles of how this portrays use in consumer settings. What is meant to be clarified here is how big data effects the process by which consumers use services and are data mined. To affectively understand this component, there needs to be an understanding that big data and it cultivation are just as it is named, a collection of data on a mass scale meant to just be data. It is however, the enterprise of the scientific community as well as commercial bodies that induce a specific plethora of uses to assimilate the necessary components to use big data effectively. Such things make it to where consumers have an easier time with their collective use of these commercial branches (Amazon and Netflix, as well as unnamed companies). Consumers use of these sites creates data, a process by which these companies then use the process of data mining to achieve the best standards of prediction and analytics that help them to enforce their use of big data as their tool in apprehending consumers by prediction. Because of this process, consumers are the ones who are helping these companies further develop their own uses of big data by allowing these corporate bodies to data mine them, collect data, and analyze the data provided. But this process is crucial to the experience of the consumer, as this data helps with creating the necessary components of these corporate bodies, allowing them to further create advanced algorithms that help these corporate bodies devise the necessary recommendations and make the predictions to the habits of these consumers. That makes it easier for consumers to use these products offered by these corporate bodies.

In all, big data, with modifications from things such as data mining, data collection, machine learning, algorithms, and prediction analyst are all components which excel the use of big data (because of the necessary enterprise that it takes to stay updated with this matter) and insure the that consumers and users of big data can reach out to their own platforms easily.

#### 4 PRODUCT RECOMMENDER SYSTEM

In the recent past, Amazon has moved from operating as a pure e-commerce firm to a major player in the internet services industry, with focus on offering a wide variety of services to both individuals as well as companies. The firm started to shift its focus on big data and started the journey to transition from a typical online retailer into one a major force in the realm of big data. Around 2000, the company, along with other internet firms such as Google, Yahoo, and Twitter realized that they had voluminous data about their customers, which could be put used to improve their performance. Although the other firms did not initially concentrate majorly on big data, Amazon swiftly moved to take advantage of the invaluable database of individuals who used its e-commerce platforms around the world to shop. The team charged with the responsibility of recommending the products to the customers came up with innovative strategies that the firm could make use of the data collected

by the firm about their customers. The end result of the move was a huge success in big data, which revolutionized how the company did business.

As a major player in the e-commerce domain, the success of Amazon was always pegged on availing the right products to the customers. The efficacy of providing the right products for the customers in turn largely depended on a proper understanding of the needs of the consumers. A proper market research was necessary in order to understand the customer's needs and tastes. Since it was founded, Amazon has created a name for itself because of its superior product recommender system, which suggests products to consumers on the basis of their last purchase. The major driving force behind the recommender system is the data gathered from the customers.

The product recommender system is essential for the personalization of each customer's experience when they are shopping in the firm's online store [25]. The firm employs collaborative filtering and clustering algorithms to classify clients on the basis of preferences. Customers are grouped on the basis of same search as well as collaborative filtering between items. Content-based search employs the shopping history of customers and item ratings to establish a search query capable of finding other items that match the tastes of consumers. For instance, if a customer purchases a book, the product recommender systems will suggest books from the same author, publisher, or subject area. The product recommendations are not only used by the company in the online stores, but it also doubles up as a marketing tool useful in conducting email campaigns. There is a recommendation link that enables shoppers to filter products by several criteria depending on the items that they have in their shopping carts.

#### 5 BIG DATA FOR DYNAMIC PRICING

Dynamic pricing entails the use of big data such as clickstreams, purchase history, cookies, etc. to offer customized discounts to customers or to alter the prices of items being sold dynamically. The technology enables the real-time price customization for an item to suit a specific customer. This explains why it is sometimes possible for two different sets of customers to buy the same item at different prices from the same online store [23]. Despite the immense benefits of this technology, some customers may always feel discriminated against due to the price differences. Amazon has successfully used the power of big data to implement a price discrimination system. For example, there was an incident in which some Amazon customers were aggravated about price variations of a certain DVD. One of the customers noted that there was a difference of nearly two points five dollars in the price if the cookies were deleted from the computer. Price discrimination was also experienced in the sale of a product known as Diamond Rio MP3 Player.

Big data also enables price optimization. This enables the firm to manage the prices of commodities and grow its profits by twenty-five percent annually. Several factors are used to set the prices of commodities. Some of them are: activity of the customer on the

firm's shopping portal, availability of the product, competitor's prices, order history, item preferences, and the anticipated profit margin [23]. The prices are normally refreshed every ten minutes as big data become updated. Due to this, Amazon provides customers with discounts on best-selling commodities and accrue large profit margins on the items that are less popular with customers.

## 6 BIG DATA AND CUSTOMER SERVICE

Big data is also extensively being used for customer service at Amazon. The acquisition of Zappos has often been viewed as a major element in the same. Since it was founded, Zappos has enjoyed a good reputation for the excellence in customer service and was usually viewed as a world leader in this domain. Much of the success can be attributed to their advanced relationship management systems which extensively employed their own customer data. After the acquisition of the firm in 2009, the procedures were integrated together with those of Amazon. Today's business environment is changing at a rapid rate, and consumers are also using their voices faster. Within a few moments after undergoing a bad experience, customers can swiftly move into social media and spread the news about their negative experience [17]. The only strategy for an organization to survive under such conditions is to employ the power of analytic to streamline and shorten the response time, as well as fix the customer support issues. The customers of the present day are not only looking for a product that works, but also one that is personalized and able to recognize their interests and save them time.

## 7 ONE CLICK ORDERING

Amazon used big data to create one-click ordering. This feature is activated automatically when the customer places his first order, enters a shipping address as well as a method of payment. When using the one-click feature, the customer is given thirty minutes to change his mind about the particular purchase. This system was created on the premise that a simplified path to purchase would increase conversion rates. Since the introduction of the technology, the firm's revenues have increased year after year. The significance of this application pushed the company to patent it to prevent other companies from using it without authorization. Reorganizing the purchase process is currently one of the most significant differentiates in the current marketplace. The service enables users to make payments without having to exchange cards or money physically. Amazon has also greatly benefited from impulse buying, which is accelerated by one-click buying. Research has shown that the largest percentage of people normally purchase things they don't require or did not plan to purchase in the first place [5].

## 8 USING BIG DATA TO SUPPORT OTHER COMPANIES

Amazon also uses its big data platform to support and help other companies improve their operations. Organizations can employ AWS toolkit provided by Amazon to create scalable big data applications that have the capacity to improve business performance [25]. Besides, they would be able to secure these applications easily without the need to spend on expensive infrastructure and hardware. The big data applications including data warehousing, clickstream

analytic, fraud detection, internet of things, and several others are delivered via cloud computing. Hence, there is no need for an organization to incur additional costs in setting up a data center. The Amazon web services can enable companies to analyze spending habits, customer demographics, and other related information to enable them effectively cross-sell some of the firm's products in patterns similar to Amazon. That is to say that the retailers will also be able to stalk their customers, recommend products to them, and improve their customer experience.

## 9 BIG DATA TECHNOLOGIES

**Amazon EMR:** This technology offers a managed Hadoop framework that simplifies and hastens the processing of huge amounts of data across scalable Amazon EC2 instances. Amazon EMR also supports other common distributed frameworks including HBase, Apache Spark, Flink, and Presto [3]. Besides, it reliably and safely handles a wide range of big data use cases, such as web indexing, log analysis, financial analysis, machine learning, and bioinformatics.

**Amazon Athena:** It denotes an interactive query service that simplifies data analysis in Amazon S3 via standard SQL. Since it is service less, one only pays for the queries they run and there is no infrastructure to be managed [3]. The technology is quite straightforward and delivers results within the shortest time possible. Moreover, it does not require complex ETL jobs to prepare data for analysis.

**Amazon Kinesis Firehouse:** This is one of the simplest methods to import streaming data into Amazon Web Services. The technology can be used to gather, transform, and import streaming data into Amazon S3, Amazon Kinesis analytic, and Amazon Redshift, to permit instant analytic with the current BI tools and dashboards currently being used. It is a comprehensively managed service that can expand automatically with the increase in data throughput.

## 10 UNSTRUCTURED DATA AND AI COMPONENTS OF SOCIAL MEDIA

Next, we will look to observe big data and its uses within social media. First and foremost, it is important to understand that social media is an outlet that is massive. The many posts, tweets, likes, shares, and other social media actions all develop unstructured forms of data that are then considered by corporations to understand the market of users. It is within the creation of this unstructured data that creates such an importance to talking about social media and its perfect relationship to big data. Because of the importance of social media to businesses (due to trend and the fast-paced living environment we live in, if a business is behind on trend it becomes behind on sales and other forms of innovations), there is a large component of dependence towards retrieving big data from social media to calculate and predict trend. But beyond just trend, businesses are also trying to get their hands on the enormous amount of big data that exists within the social media sphere. There is recognition of the value of unstructured data that is sourced within social media. The value comes from consumers using social media to broadcast what they are thinking, want, and are doing. These

types of information, one might think it private, but the internet is a very transparent source of data. In so, businesses value the perspective of consumers and create ways in which they can follow through with interacting on a very contingent basis, they data mine and from that, they advertise based on the collected big data from social media [12].

This brings to focus the use of advertisement and the necessity for big data to be used. Within digital advertising, the one who collects and analyzes big data effectively and efficiently with accurate uses is king. To be successful in this method of advertising, businesses need to be prodigious at their collection of data, integration of that data and analysis of that data. The reason being is if these three things are managed well, the use of the data is much more successful. What is challenging about this type of work however is that a majority of the data that is collected from social media is unstructured, it is in a word, messy. These forms of unstructured data are in our own use of social media, usually within posts, videos, tweets, photo post to Instagram, mass use of Snapchat, etc. Because these forms of data are so much more unstructured and more difficult to analyze using traditional analysis methods, businesses needed to enterprise methods for them to collect these data forms and have the necessary big analytics platforms to analyze the data. Keep in mind the data has the most information about us, therefore the use of these unstructured data is key to devising targeted and precise marketing executions [20].

There is also the collection of real-time data. Because of the advances within the technology, the process by which real-time data can be analyzed has sped up exponentially compared to the past where this process would be impossible and yield no results. With the ability to now look real-time data and have the capacity to analyze it, businesses (those that are marketers) have the possibility of taking action instantly to provide consumers on social media with their own personalized ad. The use of personalized ads of course comes from the massive amounts of data that consumers put out on social media, allowing marketers to collect these data (things we like, talk about, and do daily). And because they have these data, they can target specific ads to consumers without missing a beat. But what happens when we begin to incorporate other technologies that allow us to always be able to access our social medias? Mobile devices provide the quintessential provisions needed for data to have a constant flow to advertisers. Big data then is a facet within the life of social media. It must be done effectively in order to continue its uses of data collected and then expounded on to get customers to buy products or even interact with specific businesses. With the development of the mobile device, location also becomes part of big data, as it allows for your location to be collected, you leave not only a digital footprint, but also a physical one which can be collected as data and used [20].

Because of such enterprise, big data can gather almost every facet of information that is readily available to the internet. Such a mass look on data also comes back to the revision on algorithms that are meant to specify what is being observed and analyzed in the data. Of course, for these algorithms to work, there has to be a steady stream of data in order for the algorithms to process and do

its job. Because of the accessibility of data through the means of mass social media and how quickly consumers use and are exposed to social media (due to mobile phone), businesses are more creative in their approach to their algorithms. These algorithms can now pop up wherever consumers are on the social sphere. In doing this, big data itself changed the advertising platform. Advertiser now must create enticing messages from big data to continue their reach to consumers. The change is incredibly eye opening. As the use of data itself can create a massive alteration in how a marketer begins to try and craft their ads to highlight what it is that individuals want. Advertising becomes more individualized and work closely around the sphere of social media to allocate their ads effectively [9].

Within the use of big data, there is also the use of AI to help with the process of analyzing big data. AI creates a much more effective measure when it comes to analyzing hundreds and thousands of data in detail. Because AI can do that, it allows for businesses to have a better idea around what perspective they themselves must take when advertising based on detail production from AI technology. AI technology becomes a very important component to being able to go analyze big data in order to provide the necessary specific information that would help with creating the necessary ads [12].

We are then looking to see how this affects the user, or better how this type of big data in social media affects consumer experiences. The use of big data comes back to the work of the consumer and the business. The consumer creates big data from their usage of social media, social media in turn collects and responds to the data that is being created by the consumer (user). Through the process of facilitating the data, analyzing it, interpreting it, pooling it, and filtering it through algorithms and AI technology, consumers using social media get access to tailored advertisements. The consumer experiences a personalized social media experience and personalized advertisement trail based on the collected data that is reinforced by big data that is pooled together in order to do this. Does big data have any other uses in social media then? Yes, it does.

Social media can also use big data in order to create studies and infiltrate certain components of a consumer's private life. Facebook for example, a top social media company prides itself in its technological advances that use big data. Facebook uses is considered a top user of digital advertisement, so it begs to question what else does Facebook do with the mass big data that it collects? Facebook has also used its big data resources to try and act on social media experiments. During a time when there were the I Voted stickers sprawled on Facebook for users to share and gimmick that they had voted, Facebook was using this in order to incite and boost voter turnout. The method was to first isolate and use the stickers with particular groups of people, small groups at first in order to test the role. After having enough data collected on the presence of the stickers and what they meant for users, Facebook began to mass data span by incorporating the stickers in a much more massive turnout. With midterms of 2010, there is studied on behalf of Facebook scientist, that say 340,000 more people voted in the 2010 midterms [18].

From these the observation of these big data uses, we can showcase the how and what makes big data so important to the creation of diversified consumer experiences but also showcase the necessary components to how big data is used by the new technologies we have. It is however imperative that we understand these different phenomena that come from the use of big data.

## 11 SOCIAL MEDIA IS SIGNIFICANT FOR COMPANIES AND INDIVIVUALS

Although big data is said to come from several different sources, the largest proportion of it is said to originate from unstructured sources. As it can be imagined, social media makes up the largest source of unstructured content for big data. All the activities that users perform on social media such as views, retweets, comments, favorites, likes, etc. can be gathered and explored by interested individuals.

In the current digital world, social media plays a vital role in many companies. Having a presence on various social media platforms such as Instagram, Facebook, and Twitter is imperative since it enables individuals to interact with an organization on an ostensibly personal level and at the same time helps businesses across several domains get in touch with their customers. Currently, Facebook alone has over two billion users on their platform; this is roughly twenty-six percent of the world population [12]. It is therefore important to consider the fact that big data, from the social media platforms, can reach any people in different forms. Besides that, social media interactions have continued to play a big role and will continue to play a big role in business decisions. For example, some insurance companies have declined to offer life insurance policies to individuals solely based on their social media posts. If you frequently post, on any of these platforms, about how you are drinking or going to drink, insurance companies would be reluctant to offer you a life insurance policy as this is a risk to them.

It will not be long before organizations discover new and better strategies for making sense of big data. But, at the moment, the concept of big data is still new and rapidly evolving. Nevertheless, some businesses have found ways of interacting and using this data, which is just but the beginning, but still a good way to begin. To elaborate, a marketing company whose interest is promoting a new product could employ machine learning algorithms that enable it to gather data from individuals who meet certain attributes [12]. Consequently, by employing artificial intelligence technology, they will also be capable of drawing insights from millions of users and create campaigns. This will increase their levels of precision and focus, a technique usually referred to as targeted marketing, and present an excellent opportunity for finding the perfect audience and satisfy its preferences.

## 12 BIG DATA IN SOCIAL MEDIA ADVERTISING

Fundamentally, advertising revolves around communication since it is all about sensitizing consumers on products and services that an organization is selling. However, different consumers will always want to hear varied messages, which is a vital fact to consider when

new clients are being recruited into the internet bandwagon due to the growing popularity of smart phones. Big data has the capacity to customize these messages, project what consumers would like to hear, and establish new perceptions on what customers like or prefer [8]. The above steps are all revolutionary and are expected to have a significant impact on how marketers in various organizations advertise.

Furthermore, there are some occurrences which several people do not view as advertising but are still interactions between big data and marketing like product recommendation. An obvious example is Netflix [9]. Although the company does not have a concrete advertisement plan, it employs a lot of algorithms to recommend various movies and shows to its customers. The approach saves the organization a lot of money by reducing the rate of customer exit and ensures that the right shows are marketed to the right individuals. The company's strategy is to target consumers with shows specifically tailored for them. Apart from them, other firms such as Amazon, YouTube, etc. also do the same by using product recommendation to target their customers [9]. In order to stay up to date, the algorithms need constant flow of data to help it work more efficiently. With the growth of the internet, users leave huge volumes of data not only on social media platforms but also on other places they visit online in the form of a digital footprint. This provides advertisers with new avenues to tailor their messages to meet their customer demands.

The digital footprints left by online advertisers provides new insights to marketers on what a consumer really needs, and this sometimes may be more accurate than what the customer actually says on social media. However, marketers are worried about how to safeguard the privacy and security of their consumers and therefore companies that are careless in handling data collected from consumers usually ignite a backlash which greatly impact their business. Even though targeted advertising has been in existence for quite a while [9] the more the data that is collected by advertisers, the more personalized and effective marketing is expected to be. Organizations will strive not just to gather as much data as they can, but also to gather information which typically represents the individual consumer's needs in order to enable them to market to their personalized tastes.

## 13 ANALYZING LINKS

Big data collected from social media can lead to the discovery of new information regarding each individual customer that can help in creating a customized appeal to that specific customer. However, with the new insights, marketers can enhance how advertising is approached as they create new strategies. The new growth in content marketing is usually perceived as a primary beneficiary of big data, although the concept of content marketing could be older than the internet itself.

Another essential point is that big data enables digital marketers to target users effectively with more personalized advertisements which they might prefer to see. Facebook and Google are among the biggest players in this domain of digital advertising. They have

discovered excellent ways of creating and delivering more appealing advertisements in ways that do not intrude on the rights and preferences of the consumers [10]. Most of their advertisements feature services and goods that consumers would like most to enhance their lives and almost all of these advertisements are reliant on huge amounts of personal data that users usually provide from what they are up to, what they share and like things online.

Experts contend that it is possible to accurately make predictions on an array of individual attributes that are more sensitive merely through an analysis of an individual's Facebook or Twitter likes [20]. For example, the likes on these social media websites are critical in predicting one's religion, sexual orientation, emotional stability, life satisfaction, age, relationship status, and many other attributes. Companies like Facebook successfully linked political activity with user commitment when they created a sticker enabling most of their users to declare on their profiles that they had voted. The initiative was conducted during the 2010 midterm polls and was very effective as more people turned up to vote as compared to the 2006 midterm elections [18]. Individuals who saw the feature had high chances of voting and actively engaged in a conversation about the same after seeing their friends and peers participate in the activity. Later on, during the 2016 polls, Facebook escalated their role into the voting process by providing users with not only constant reminders but also with directions about their polling stations [19]. Apart from that, they also enabled users to easily get access to registration information, news, voting guides and other tools that would have made them more equipped to go through the election process.

## 14 USER RATING AND POP UP ADS

Depending on the user preferences and the content that they often access on social media, pop-up advertisements can be created to target users every time they are online. For example, an ad can be created on the Facebook Messenger app to open inside that particular app every time the user hits the CTA button. When clicked, such ads would redirect the user to a page where they would be required to answer some question, claim a reward or send some feedback regarding a product or service. Before creating such ads, it is imperative to establish a custom audience of the individuals who would be targeted with that particular pop-up ads. For instance, individuals who have previously liked the company's products on their Facebook page or other social media sites can be included on the list of target audience to receive the ad [4]. Another strategy that can be employed is to rate users by tracking their cookies. In most cases, user activities are usually tracked across the internet using cookies whenever a user logs into one of the social media sites and is concurrently browsing other sites. Whenever this happens the other sites that the user is visiting can be easily tracked and the data used accordingly.

## 15 RELEVANCY OF BIG DATA ANALYTICS IN GROCERIES STORES

### 15.1 Increases the customer shopping experience

As per a current SHSFoodThink white paper "Are We Chain Obsessed?" 64% of customers said that the previous shopping experience is what makes them keep coming back! not the items themselves [24]. By utilizing bits of knowledge received from the information transaction database, online networking, promotional activity, customers purchasing behavior, and client movement patterns, grocery stores can find a way to guarantee they are engaged with their customers that matter most.

For instance, they can investigate customers shopping movement to enhance the layout of their store, or recognize attrition risks for clients who have not as of late bought staple things, similar to milk. In like manner, chains can construct item varieties demonstrated with the customer needs and purchase patterns in certain regions [2, 13, 24]. Regardless of whether it is through reconsidering store layout or furnishing store attended with mobile apps to better serve clients, analytics can enable grocers to change consumer's expectations.

### 15.2 RESTRUCTURE THE SUPPLY CHAIN

Grocery stores can likewise utilize analytic to investigate the production of their products, monitor production processes, and quality control, and improve straightforwardness with buyers about their sustenance production practices of foods [16]. Suppliers remain to profit from the evaluation also, with access to secure, customized content of information identified with performance sales of the product, stock, margins, and marketing effectiveness. Giving supplier an opportune profitable business knowledge that supports joint ventures, drives performance, and decreases waste products

### 15.3 BUILD SUPERIOR MARKETING PROGRAMS

Loyalty programs furnish grocery merchants with an abundance of data to enable them to distinguish client segments and precisely characterize item preferences. By joining this information with different data sources! like healthful patterns, favored technique for accepting marketing promotion, customer movement patterns, and weather-related event! grocery merchants can concentrate on enhancing, and derive income from, the general shopping experience [24]. For instance, grocery retailers can utilize analytics to customize the advancements they offer to clients given what they are well on the way to buy. They can likewise time advancements fittingly, and offer codes to customers who often as possible buy certain things.

### 15.4 IMPROVES HR STRATEGIES

Supermarket stores utilize analytics to manage work-related decisions. Information freely accessible through online networking accounts and different means can be examined in conjunction with a grocer's internal information to direct decision identified with selection and recruitment, employee termination, and performance management and advancements [11]. For example, an investigation

of late action on LinkedIn can reveal insight into which representatives are destined to leave an organization.

Grocery merchants can likewise break down information to control the advancement approaches that will build workforce performance. For example, they could explore different avenues regarding organizing a social gathering for representatives at a subset of their stores, and analyze information on profitability, morale, and turnover in the preceding months [13]. They may find that the gathering information prompted a more positive workplace where workers feel more noteworthy engagement at work, and soon after that, they could roll the strategy out to different stores.

## 15.5 USING BIG DATA FOR COMPETITIVE ADVANTAGE AND ATTRACTING CUSTOMERS

Numerous grocery stores have been utilizing transaction and client information for a considerable length of time, despite the fact that many still have not completely used all that can be proficient with these types of information. For Small to Medium Sized grocery merchants, many have swung to subcontracted point solutions because of an absence of available analytics assets and potential framework investment required [11, 24]. The issue with point solutions recently is that if? they independently work out for a particular business section and the evaluation is cookie cutter. In this way, the 'information' is not coordinated and hard if not difficult to give an all-encompassing picture of client conduct overall touch focuses for instance. Nor are the investigations offering a cross-functional observation that is pertinent to all business partners as far as driving differentiation in the commercial center in promoting, advertising, store operations and supply chain.

As far as utilizing 'new' data sources, for example, mobile, social and text, the industry is particularly occupied with a discovery phase of investigation with an assortment of center sections, testing and figuring out how to extricate an incentive from these rich new sources of information. There are two common paths grocery merchants takes with little respect of the 'size' of the organization: to start with is Strategic Commitment, in which there is C-level (hierarchical) commitment making the venture in the assets to get the majority of the in-house data and evaluated it [13].

Presently like never before, information, analytics, and IP are seen as vital resources and competitive discriminators. The other is Business Discovery; in which grocery merchants outsource to an Analytics as a Service firm to use internal and external information. Performing analytics speeds the construction of business advantages creating new users case and helps catch 'quick wins' before making resource commitment to technological innovation and human capital in advance [11]. In view of progress, and a wit, trusted stakeholder willing to share the techniques and explanatory models, can assist grocery merchants to proceed with an outsourced administrations supplier or relocate the data, analytics in addition to IP in-house.

## 16 RECOMMENDATIONS

### 16.1 Real-time insight on product demand

Nowadays, retailers can get to information on item demand levels instantly on a chain of stores. Nevertheless, numerous merchants are still in the earliest stages in regards to evaluating and monetizing the huge amount accessible data [2]. This prompts stocking deficits, for example, evaluating item demanded based exclusively on past historical information. It can likewise convey about wrong promoting endeavors: If a customer purchased ketchup on Saturday, an email coupon for it on Sunday is not well planned and make little sense to the shopper.

This is the place data from store loyalty programs in addition to credit card sales can prove to be useful. Its data can be utilized to define needs of the customers in future. For example, grocery merchants can use data analytics to decide how regularly customers purchase sugar, flavors, or different items, and after that send every family unit coupons given their propensity to buy [24].

### 16.2 Enhancing in-store stock management

Perishable basic supplies, for example, dairy, meat, and fish call for precise stock administration, regularly on an hourly premise. Client analytics and prediction tools can enable grocery merchants to calibrate their inventory levels by assessing buyer purchasing behavior and requested products from various viewpoints and situations [24].

For example, grocery retailers might need to screen cycles like when customers go for particular nourishment, purchasing patterns amid sales deals when storing activity peaks or seasonally inspired buys. As indicated by a report from Manthan, this methodology worked for U.K. food grocery merchant Waitrose: a deeper understanding of buyer purchasing behavior and demand outlines using cutting edge client analytics and predicting tools helped the store [11]. Concurrently, retailers can utilize these systems to all the more deftly change their stock levels and amplify high-buy products.

### 16.3 Leveraging Predictive Analytics

Amazon spearheaded item proposal engine: the "if you purchased that, you may like this" invention. This strategic changing web-based shopping feature mirrors the retailer's profound assessment of buyers' shopping basket. Proposal engine is intended to enable customers to find items they were not sorting out but rather would be interested in purchasing [13]. Today, general grocery merchants are progressively tapping the global innovation behind proposal engine: predictive analytics. This kind of assessment measures future patterns in light of present and past information, and it can enable stores to improve business. Information is driven, all-encompassing assessment of "purchasing triggers, for example, regularity, weather, stock, and advancements, is progressively informing grocery stores' product blend, marketing plans, and sales forecast [2]. Furnished with these information-driven tools, stores can better distinguish what items customers need today and what they will be demanding in future, and this learning will enable them to stay competitive for a considerable length of time to come.

## **17 INTRODUCTION**

Digitization set apart by an increasing number social media and mobile devices is shifting the business landscape in every sector insurance included. The opportunity presented by this aspect for insurance companies are immense. Communities and social networks enable insurers to interface with their clients better, which to their advantage improves branding, customer retention, and acquisition [24]. Insurance companies additionally get a plenty of contributions from computerized data as feedbacks, which likewise can be utilized to develop unique products and aggressive valuing. Digitization of big data analytics offers numerous opportunities that Insurances Company can harness to detect fraud among their customers. Dealing with fraud manually has dependably been expensive for insurance firms regardless of the possibility that maybe a couple of minor fraud went undetected [6]. What's more, the trends in big data (the evolution in unstructured information) are prone to numerous fraud, which can go without notice if analysis is performed correctly. In the proceeding section, the article will examine important of big data in insurance fraud detection and its relevancy.

## **18 IMPORTANCE BIG DATA AND INSURANCE FRAUD DETECTION**

Conventionally, insurance firms utilize statistical models to recognize fraudulent cases. These models have their limitation [15]. To start with, they employ sampling techniques to assess information, which prompts at least one fraud going unnoticed. There is a punishment for not performing a proper assessment of the data provided. Subsequently, this strategy depends on the cases analyzed before. Therefore, every time different fraud takes place, insurance firms need to manage the impact for the first time. Lastly, the conventional strategy works in silos and is not correctly equipped for taking care of the natural developing wellsprings of data from various diverts and diverse capacities in an integrated way. Analytics tends to be difficult and assumes an exceptionally pivotal part in fraudulent recognition for insurance firms. In the proceeding section, the significant benefits of utilizing big analytics in fraud detection assessed.

### **18.1 Identification of low incidence events:**

Utilizing sampling methods accompanies its particular arrangement of acknowledged mistakes. By using analytics, insurance can manufacture frameworks that go through every fundamental datum. This like this distinguishes events with low frequency (0.001%) [7]. Methods such as predictive modeling can be utilized to altogether break down processes of fraud, channel clear cases, and allude low-rate fraud cases for facilitating analytics.

### **18.2 Enterprise-wide solution:**

Analytics help in building a global point of view of the anti-fraud endeavors all through the undertaking. Such a point of view regularly prompts dominant fraud location by connecting related data inside the association. Fraud can happen at various source focuses premium, claims or surrender, application, employee-related or outsider fraud. In the meantime, insurance channel broadening is

adding to the breakdown of identifiable information. Insurance-related exercises should be possible using cell phones separated from the conventional face-to-face and online Insurance [15, 25]. This can be seen as an expansion to data storehouses in the Insurance business. Given more prominent channel enhancement and the development of ranges where fraud can happen, it is vital for insurers to have reachable enterprise-level data about their business and clients.

### **18.3 Data Integration:**

Analytics assumes a vital part in incorporating information. Viable fraud recognition abilities can be worked by joining information from different sources. Analytics additionally help in integrating inside information with outsider information that may have predictive significance, for example, public records. Information sources with derogatory properties are on the whole public documents that can be incorporated into a model. Cases include liquidations, liens, criminal records, judgment, abandonment, or even deliver change speed to show transient conduct. Different sorts of outsider information can be useful in upgrading effectiveness, for example, audit evaluating data to decide whether harms coordinate portrayal or misfortune or injury being guaranteed [6]. A standout amongst the most under-used information sources is doctor's visit expense audit information. This information, if utilized as a part of a model legitimately, is a gold dig for organizations researching medical fraud. Revealing peculiarities, in charging and adding these to the next scoring motors or interpersonal organization analytics will diminish the measure of time an agent or expert spends endeavoring to pull the majority of the pieces together to recognize deceitful action.

### **18.4 Harnessing Unstructured Data:**

Analytics is useful for getting the best incentive from unstructured information. Fraud can be delicate or hard. This depends on whether it comprises of a policyholder's misrepresented cases, or on the off chance that it contains of a policyholder arranging or creating a misfortune. At an abnormal state, fraud can happen amid commission discounting, because of false documentation, an arrangement between parties or from is offering [24]. Albeit bunches of organized data is put away in an information distribution center as a component of numerous applications, a significant portion of the vital data about a fraud is in unstructured information, for example, outsider reports, which are not assessed. In most insurance firms, data accessible in online networking is not suitably stored. An uncommon investigative-unit specialist will concur that unstructured information is vital for fraud examination. Since textual information is not straightforwardly utilized for reporting, it does not discover a place in most information stockrooms [7]. This is the place content examination can assume a crucial part in checking on this unstructured information and giving some valuable experiences in fraud discovery.

## **19 RELEVANCE OF BIG DATA IN INSURANCE FRAUD DETECTION**

Big data analytics is a reality for the insurance company because of its capability to enhance various conventional technologies and

be used to detect fraudulent acts. In the proceeding section, the relevance of big data and insurance fraud detection will be examined.

### 19.1 Text analysis

In numerous Insurance fraud recognition ventures, from 33% to oneportion of factors utilized as a part of the fraud location model originate from unstructured content data. This is particularly helpful for long-tail claims, for example, damage claims, because the best information frequently is found in claim notes [15]. Content mining is something beyond keyword sorting. Excellent content analytics apparatuses translate the importance of the words to establish context. Innovation that is adroit at preparing common dialect can help remove factors from the unstructured content that can be utilized for assist fraud modeling.

### 19.2 Data Management

Regardless of where your information is stored from legacy frameworks to the valid information stockpiling structure, Hadoop an information administration framework can enable insurers to make reusable information rules. They give a standard, repeatable strategy for enhancing and incorporating information [7]. Preferably, you need a framework that interfaces with different information sources. It ought to have streamlined information league, relocation, synchronization, organization, and visual assessment.

### 19.3 Event Stream Processing

This enables insurers to investigate and processes in movement (i.e., process streams). Rather than putting away information and running questions against data, you store the inquiries and stream the data through them [24]. This is foundational to both ongoing fraud identification (invigorating fraud scoring) and successful utilization of great high-speed information sources similar to vehicle telematics.

### 19.4 Hadoop

A free programming structure that assesses and prepares of tremendous collected information in a distributed environment of computing. It offers gigantic details stockpiling and super-quick processing at around 5 percent of the cost of convection less-adaptable databases. Hadoop's mark quality is the capacity to deal with organized and unstructured information (counting sound, text, and visual), and in expansive volumes. Insurers either can employ Hadoop specialists to exploit the structure or purchase items that scaffold to existing databases and information distribution centers[6, 7]. This foundational innovation for making predictive analytics models stays one-step in front of fraudsters and spillage of paid-out cases cash. The exchange observing advancement innovation used to battle regularly complicated illegal tax avoidance utilizes Hadoop as a center stockpiling and sorting out innovation. Complex organized crack rings and therapeutic factories, for instance, are conveying progressively modern techniques for laundering cash stolen from auto insurers.

### 19.5 In memory

In-memory analytics is a processing style in which all information utilized by an application is put away inside the principal memory

of the computing condition. Instead of being available on a disc, the data stays suspended in the mind of useful sets of PCs. Different clients can share this information with numerous applications in a quick, secure, and simultaneous way. In-memory analytics likewise exploits multi-threading and distributed registry [6, 24]. This implies clients can disseminate the information (and complex workloads that process the data) over different machines in a group or inside a single server condition. In-memory analytics manages questions and information analytics, yet also is utilized with morecomplex procedures, for example, predictive analytics, machine learning, and analytics. The sorts of neural-network analytics that assist insurer in discovering association among suspects sustaining claim and premium fraud depending on the kind of processes

### 19.6 Software as a Service (SaaS)

Predictive modeling and different analytics were accessible to large insurance net providers willing to introduce the innovation on location as of not long ago. Software as a service has advanced to even where genuinely little insurers can exploit Big Data analytics [6]. Insurance providers subscribe to a service keeps running by a seller as opposed to paying for the vast buy, establishment, and support of in-house frameworks. SaaS likewise is named "on-demand software."

## 20 DISCUSSION

Form such progress we can see the diverse reach that big data has and how it affects the users in their experiences with big data. Not only do we see big data creating advertising products to consumers, we are also seeing social media sites using big data to influence other functions of consumeris daily lives. To further that observation, there is necessity behind seeing the claims that Facebook makes on its ability to influence its consumers to influence those within the social media sphere. If big data has the access to arrange itself around the sphere of the consumer to have the consumer act on certain task, what does that mean the uses of big data are to social media sites? We can assume from our knowledge that social media sites like Facebook could be interjecting into the private lives of their users by instigating on the data that is collected in processed to pinpoint user habits. It can also be seen that big data facilitates the necessary components of information to allow social media sites to specify their own approaches to their consumer in ways that can be seen as going over the line when it comes to their connection with their consumers. It is however, still a very enterprise avenue of using big data. As it allows for the social media to influence social, economical, and political landscapes. But that in itself is also a very dangerous power to have. As those who use the big data and direct their resources into specific marketing strategies can alter nearly whatever they fid like to in front of the irrational consumer.

As we have observed of big data above, we also learn of the prediction value and how big data escalates the ability for businesses to predict and recommend to consumers different products. The use of pooling data and sifting it through algorithms in order to precisely choose methods of spawning products before consumers

is a fundamental use of big data. Big data becomes a tool that assimilates the data that is created to businesses to create more big data. The process is unending and constantly provides businesses with unlimited amounts of data that can be used to spearhead their campaigns. Does big data then become a commodity that is used like currency to businesses? Well, it is very possible. As the use of big data is how businesses maneuver their strategies to get consumer to consume. If these algorithms meant to increase sales were used for something else, say medical awareness to the issues that exists within smoking cigarettes, how will the big data be used and what forms of algorithms would be used? The use of prediction analyst suits sales, but based on the observation made above, it is probably even more effective in helping to create a knowledgeable public. By facilitating the big data and being able to sort out the necessary public images that have control over social sphere through mass social medias, there can be an exchange of data between consumers. Because of the interplay between data and how consumers absorb them and create more data that is spawned for more information, big data can in turn control knowledgeable outcomes in public opinions. The use of big data is vast then when it comes to understanding the many components that make up what big data is.

Our observation above also places the consumer experience as an important facilitator for big data to exist. The habits and practices of consumers as well as the opinions and locations of the consumer can truly inhibit how big data is filtered to facilitate the necessary components of data to market and process information. This process can of course be seen using AI technology in order to expedite the data that is coming in. It does this so that the data is filtered and able to be used to immediately influence commercial markets, social media spheres, and consumer habits. That in turn regulates and begins to push out even more big data from the interactions that consumer have with the new platforms that are created from old big data that was used in order to create their new purchases or opinions. AI technology then becomes a fundamental component to the access, collection, analysis, and interpretation of big data. Its use of manipulating and translating data in order to be used to create enterprise is crucial to the development of more big data. From the observation above, it can also be said that AI technology will has also positioned itself in a way where it has become fundamental to the big data analysis and because of that, AI technology is part of big data.

The paper also sought to observe the nature of consumers having the capacity to control the big data that flows into collection. The use comes from filter settings like those included on Amazon in order to help narrow searches of items or on Netflix in order to create the right kind of streaming that the consumer requires or likes. If that is the case, the consumer actually holds a lot of power when it comes to the collection, analysis, and interpretation process of big data. Within how the consumer chooses to reside over these social medias and commercial businesses determines how the social media sites and businesses get their data. Beyond that, the consumer has no knowledge of how to analyze or interpret big data, yet holds the key to the very idea of big data. Because of this notion, the discussion here seeks to try and highlight the importance of

businesses maintaining and using data collected responsibly.

Big data is used in order to interact with consumers in order to sell or sway. The use of data however is also created by consumers. For this symbiotic relationship to exist and stay peaceful, businesses must be sure that they are not over stepping privacy issues when it comes to the use of big data. By enterprise methods to help consumers with choices and options through their recommendation systems and early predictive measures of trend by their collected big data, that is fine. But when business pressure consumers with the use of big data, the business will most likely end up losing new data to collect. As if no one is using their sources to create data, their lack of data causes them to have slow flow, and that leads to isolation of data.

Take for instance the process of data mining. Data mining is used in order to receive particular forms of information about consumers. This information is in the form of data, this data is put through special filters to narrow in on what it is that businesses want to know about particular groups of consumers to achieve the best methods of interacting with the consumers in order to highlight necessary products to the consumer. What happens if the algorithms for this particular data mine was off? This would mean that the data that was supposed to continue in the line of procession ends up lost. Because missing the mark with data mining and interpretation means that businesses loses their edge with their consumers.

Big data is a very complex topic to talk about. It is however, a very interesting topic to look at. As when we are observing what forms of big data are used in order to create experiences for consumers and business practices for businesses we can see the importance of having a very strong handle on the idea of big data. It is not just a process by which you collect massive amounts of information and then through it back out into a market. Big data must be molded around using algorithms, AI technology, studies done to mine particular forms of data, and even understanding the complex notions of unstructured data. Because of these reasons, the study of big data is still relatively incomplete. The use of big data however, should be understood as a relationship between consumers and those who seek to use big data to facilitate their individual means.

## 21 CONCLUSION

The paper sought out to examine the complexities of big data, but to be more precise, this paper seeks out to the multiplicity of avenues in which big data is used by our technological world in regards to Online shopping, Streaming Services, and Social Medias. The conclusion is that the multiple complex systems which make up the forefront and system of enterprise around big data falls under the very distinctive relationship that exists around the users of these services, and the sources the users use in order to create more big data.

Such complex multiplicity of diversified uses alter the understanding of big data by showcasing that big data in itself is easily

manipulated and altered. This becomes the case because of the multiple layers of data that exists in any given moment. The use of these data are incorporated in a way that there are many organization that are still trying to spearhead further in the endeavors of big data. The papers observation of the multiple forms of big data conversions, analytics, prediction standards, and even experimental uses of data reinforces the concept that big data is as it is called, massive. Because of such presence, to truly be able to look into the multiplicity of big data would mean a massive overhaul of research meant to showcase the existence. This paper however does not do that, but would also rather seek to showcase that.

Online shopping and streaming sites uses a multiplicity of tools alongside big data in order to function with consumers and creates a diversified experience for consumers. The tools that are used by these shopping and steaming businesses alter big data into sustainable forms of information that are then used in order to predict and recommend to consumers what products to purchase and recommendations. These are all done through algorithms used to analyze the data. The paper observes and concludes that the standing made by these businesses are diversified and are meant to showcase the suitable substance of the big data to consumers. The use their procedures not only diversify the consumer experience but also diversifies the way that big data is collected and used. Big data within these forms of principles are collected and retained under algorithmic databases that are then filtered out when it is being generated by consumers. The process of big data filtering is not only done by businesses, they are also given as options to consumers. Through the process of filtering consumers can do the exact same thing.

Social media sites use big data just at online shopping and streaming services do, but social media also has another power. One which allows them to facilitate their studies into experiences around the consumer. By doing this, social media can use big data in order to influence and manipulate consumers into specific acts of studies that are being done by the social media sites. This form of big data usage not only diversifies but also includes possibility for growth in collection of information. As when social media analysis these complex forms of data known as unstructured data, they are having a deeper perspective into consumer habits, wants, and additional personal information that expands their usage of big data.

Big data is a very diversified entity. Even though it can be narrowed down to certain institutions or entities, it in itself is able to expand largely in those narrowed views. The diversification of big data is very crucial to the survival and usage of big data. Without multiple sources to collect necessary data, there would be no big data. Therefore, the userfis experiences around big data needs to be one that is flexible and seeks to incorporate the right amounts of AI and Algorithms in order to maintain steady flow of data which encapsulates the very idea of diversified big data.

In summary, the multiplicity of avenues that exists around big data creates the very core study that it takes in order to understand the practices and exhibits which are induced by the use of big data itself. By observing real world applications of big data we can see

that the diversification of big data does not need to be mellowed or shallowed out from the perspective of a programmer, as it seems that the market capitalizes on big data and therefore creates the very enterprise of multiplicity within its diversification.

## ACKNOWLEDGMENTS

My group work very hard to facilitate a study around diversification and observation. Their hard work in reading through these sources multiple times to see the rights of the observation is very much appreciated. We would also like to thank our professor for the chance to take on such a free ranging topic, as it has allowed us to further appreciate big data and its importance in our future endeavors.

## REFERENCES

- [1] Bertino, Davidson S. Dayal U. Franklin M. Agrawal D., Bernstein P. and Colleagues. 2012. Challenges and Opportunities with Big Data: A White Paper Prepared for the Computing Community Consortium committee of the Computing Research Association. (2012). <http://cra.org/ccc/resources/ccc-led-whitepapers/>
- [2] J. Aloysi, H. Hoehle, S. Goodarzi, and V. Venkatesh. 2016. Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. (2016). [http://www.vvenkatesh.com/wp-content/uploads/dlm\\_uploads/2016/07/2016-AOR-Aloysi-etal.pdf](http://www.vvenkatesh.com/wp-content/uploads/dlm_uploads/2016/07/2016-AOR-Aloysi-etal.pdf)
- [3] Amazon. 2017. Big Data on AWS. (2017). <https://aws.amazon.com/big-data/>
- [4] John Aycock. 2010. Springer Sciences & Business Media. *Spyware and Adware* 50 (2010), 71–109.
- [5] Roy F Baumeister. 2002. Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of consumer Research* 52, 4 (2002), 670–676.
- [6] Chui M. Brown, B. and J Manyika. 2011. Are you ready for the era of fibig datafi? *McKinsey Quarterly* 4, 1 (2011), 24–35.
- [7] A. A. Crdenas, P. K. Manadhata, and S. P. Rajan. 2013. Big Data Analytics for Security. *IEEE Security Privacy* 11, 6 (2013), 74–76.
- [8] Kyle Hensel & Michael H. Deis. 2010. Using Soical Media To Increase Advertising And Improve Marketing. *The Entrepreneurial Executive* 15 (2010), 87.
- [9] Gary Eastwood. 2017. Big Data, Algorithms and the Future of Advertising. (2017). <https://www.networkworld.com/article/3194585/big-data/big-data-algorithms-and-the-future-of-advertising.html>
- [10] W. Glynn Mangold & David J. Faulds. 2009. Social media: The new hybrid element of the promotion mix. *Business Horizons* 52 (2009), 357–365.
- [11] Ban G-Y. 2014. Business analytics in the age of big data. *Business Strategy Review* 25, 3 (2014), 8–9.
- [12] David Geer. 2017. Will Big Data Change how we use Social Media? (2017). [https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/#.tnw\\_DPCeKg97](https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/#.tnw_DPCeKg97)
- [13] M. Ghemsoune, M. Lebbah, and H. Azzag. 2016. State-of-the-art on clustering data streams. *Big Data Analytics* 1, 1 (2016), 134–145.
- [14] Amir Gandomi & Murtaza Haider. 2015. Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management* 35, 2 (2015), 137–144.
- [15] Shaun Hipgrave. 2013. Smarter fraud investigations with big data analytics. *Network Security* 13, 12 (2013), 7–9.
- [16] A. Hussain and A. Roy. 2016. The emerging era of Big Data Analytics. *Big Data Analytics* 1, 1 (2016), 249.
- [17] Randal E. Bryant & Randy H. Katz & Edward D. Lazowska. 2008. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. (2008). <https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big>Data.pdf>
- [18] Dara Lind. 2014. Facebookfis fil Votedfi Sticker was a secret experiment on its users. (2014). <https://www.vox.com/2014/11/4/7154641/midterm-elections-2014-voted-facebook-friends-vote-polls>
- [19] Sarah Perez. 2016. Facebook gives its Election 2016 hub top billing by pinning it to your Favorites. (2016). <https://www.qubole.com/blog/big-data-advertising-case-study/>
- [20] Nate Philip. 2014. The Impact of Big Data on the Digital Advertising Industry. (2014). <https://www.qubole.com/blog/big-data-advertising-case-study/>
- [21] Randy H. Katz Randal E. Bryant and Edward D. Lazowska. 2008. Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society. (2008). <https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big>Data.pdf>
- [22] Chetan Sharma. 2014. Big Data Analytics Using Neural Networks. (2014). <http://scholarworks.sjsu.edu/etd.projects/368>

- [23] Benjamin Reed Shiller. 2014. First-Degree Price Discrimination Using Big Data. (2014). [http://benjaminshiller.com/images/First\\_Degree\\_PD\\_Using\\_Big\\_Data.Jan.18.\\_2014.pdf](http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data.Jan.18._2014.pdf)
- [24] Eric Siegel. 2013. *Predictive analytics: the power to predict who will click, buy, lie, or die*. Vol. 51. Wiley, New York.
- [25] Hsinchun Chen & Roger H L Chiang & Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012), 1165–1188.

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16 while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., while executing---line 3229 of file ACM-Reference-Format.bst

```
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 48 error messages)
make[2]: *** [bibtex] Error 2
```

latex report

[2017-12-16 09.32.01] pdflatex report.tex



```
Missing character: ""

Typesetting of "report.tex" completed in 1.3s.
./README.yml
 56:72    error    trailing spaces  (trailing-spaces)
 57:75    error    trailing spaces  (trailing-spaces)
 58:75    error    trailing spaces  (trailing-spaces)
 59:77    error    trailing spaces  (trailing-spaces)
 60:77    error    trailing spaces  (trailing-spaces)
 61:75    error    trailing spaces  (trailing-spaces)
 62:77    error    trailing spaces  (trailing-spaces)
 63:78    error    trailing spaces  (trailing-spaces)
 64:77    error    trailing spaces  (trailing-spaces)
 65:77    error    trailing spaces  (trailing-spaces)
 66:78    error    trailing spaces  (trailing-spaces)
 67:76    error    trailing spaces  (trailing-spaces)
 68:51    error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Shiqi Shen
hid: 109
paper1: complete 100% Oct 27th
paper2: complete 100% Nov 4th
project: 100% Dec 4th
```

```
yamlcheck
```

---

```
wordcount
```

---

```
13
wc 109 project 13 11811 report.tex
wc 109 project 13 11842 report.pdf
wc 109 project 13 891 report.bib
```

find "

---

- 160: As per a current SHSFoodThink white paper "Are We Chain Obsessed?" 64{\%} of customers said that the previous shopping experience is what makes them keep coming backnot the items themselves \cite{12}. By utilizing bits of knowledge received from the information transaction database, online networking, promotional activity, customers purchasing behavior, and client movement patterns, grocery stores can find a way to guarantee they are engaged with their customers that matter most.
- 208: Amazon spearheaded item proposal engine: the "if you purchased that, you may like this" invention. This strategic changing web-based shopping feature mirrors the retailer's profound assessment of buyers' shopping basket. Proposal engine is intended to enable customers to find items they were not sorting out but rather would be interested in purchasing \cite{10}. Today, general grocery merchants are progressively tapping the global innovation behind proposal engine: predictive analytics. This kind of assessment measures future patterns in light of present and past information, and it can enable stores to improve business. Information is driven, all-encompassing assessment of "purchasing triggers, for example, regularity, weather, stock, and advancements, is progressively informing grocery stores' product blend, marketing plans, and sales forecast \cite{14}. Furnished with these information-driven tools, stores can better distinguish what items customers need today and what they will be demanding in future, and this learning will enable them to stay competitive for a considerable length of time to come.
- 260: Predictive modeling and different analytics were accessible to large insurance net providers willing to introduce the innovation on location as of not long ago. Software as a service has advanced to even where genuinely little insurers can exploit Big Data analytics \cite{16}. Insurance providers subscribe to a service keeps running by a seller as opposed to paying for the vast buy, establishment, and support of in-house frameworks. SaaS likewise is named "on-demand software."

passed: False

find footnote

---

passed: True

```
find input{format/i523}
```

---

```
5: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

WARNING: algorithm and above may be used improperly

70: As a programmer, machine learning and the creation of these algorithms truly fascinate when they are being used in order to create a recommendation system. Since the collection of data is pooled, the use of algorithms to reach a particular end, based on data analyzation, makes for a very interesting usage. As the use of data algorithms not only seek to facilitate the necessary recommendations, but also filter through millions of data informatics to agitate the necessary products to ensure the transparency of recommendations created. This can also be seen in what was stated above regarding consumers being able to filter products with precise search options. These filters can act as filters for filters. As in the filters chosen by consumers help to facilitate precise sifters and allows for algorithms to pinpoint precise outlines of data that help to recommend even finer tuned recommendations for consumers. \\

WARNING: algorithm and above may be used improperly

266: As we have observed of big data above, we also learn of the prediction value and how big data escalates the ability for businesses to predict and recommend to consumers different products. The use of pooling data and sifting it through algorithms in order to precisely choose methods of spawning products before consumers is a fundamental use of big data. Big data becomes a tool that assimilates the data that is created to businesses to create more big data. The process is unending and constantly provides businesses with unlimited amounts of data that can be used to spearhead their campaigns. Does big data then become a commodity that is used like currency to businesses? Well, it is very possible. As the use of big data is how businesses maneuver their strategies to get consumer to consume. If these algorithms meant to increase sales were used for something else, say medical awareness to the issues that exists within smoking cigarettes, how will the big data be used and what forms of algorithms would be used? The use of prediction analyst suits sales, but based on the observation made above, it is probably even more effective in helping to create a knowledgeable public. By facilitating the big data and being able to sort out the necessary public images that have control over social sphere through mass social medias, there can be an exchange of data between consumers. Because of the interplay between data and how consumers absorb them and create more data that is spawned for more information, big data can in turn control knowledgeable outcomes in public opinions. The use of big data is vast then

when it comes to understanding the many components that make up what big data is.\\

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

Database file #1: report.bib

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." has a comma at the end while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and O'Neil C." while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U., Franklin M., and C  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Agrawal D., Bernstein P., Bertino., Davidson S., Dayal U.,  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Brown, B., Chui, M. and Manyika, J" for entry 16  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska" has a comma at the e  
while executing---line 3229 of file ACM-Reference-Format.bst  
(There were 48 error messages)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 8217  
non ascii found 8217  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# Big Data Analytics on Food Products Around the World

Karthik Vegi

Indiana University Bloomington  
College Mall Apartments  
Bloomington, Indiana 47401  
kvegi@iu.com

Nisha Chandwani

Indiana University Bloomington  
Park Doral Apartments  
Bloomington, Indiana 47408  
nchandwa@iu.edu

## ABSTRACT

Food is one of the basic necessities of human-being. It helps us gain energy to recharge our body to do the daily activities of moving, playing, and thinking. From being a cave man to producing a wide variety of foods, we have come a long way. The civilizations shaped the food habits of the world and there is a lot of variance in the food habits across countries. We analyze the *Open Food Facts* database that gathers information on food products from around the world to unearth some food habits of the world and we predict the food grade based on the nutrition facts of the food products.

## KEYWORDS

i523, hid231, hid203, big data, food habits, food products, nutrition

## 1 INTRODUCTION

*Open Food Facts* is a non-profit initiative started by Stephane Gigaandet and run by thousands of volunteers around the world. Any person around the world can contribute to the database by simply scanning a product using a mobile app which is made available to IOS and Android. This massive database of food products opens up a lot of opportunities to analyze the food products around the world and understand the food habits. We are particularly interested in the consumption of nutrients that come along with the food items across the world, the composition of different fat content, and the prediction of nutrition grade based on the nutrients.

## 2 FOOD ANALYSIS: IMPORTANCE AND RELATED WORK

In recent times, more and more companies try to market their food as low-fat or low-calories in order to fool consumers into buying their products. The increasing concern of public health has led to a significant interest in detecting the health-related properties of food products [2]. Thus, there is no question about the importance of analysis of the nutrition grade and food safety in today's world. The analysis of food requires more robust and efficient methodologies in order to ensure the quality and safety of the food products [2]. Previous methods based on the so-called wet-chemistry have now evolved into more powerful techniques which are used in the food laboratories. These methods provide a massive improvement in analytical accuracy thus expanding the limits of food applications [2]. The traditional methods of food analysis can be classified based on the underlying principle. Some of these categories are spectroscopic, biological, electrochemical, supercritical fluid chromatography [2]. All these techniques provide information about the sample under study and this information is derived from a specific physical-chemical interaction [2]. A different approach to analyzing and detecting the food quality is by using machine

learning techniques. We will discuss one of these modern methods of food analysis which can be widely used across countries.

## 3 ANALYSIS OF NUTRIENTS IN FOOD

Fat is definitely a nutrient that the body needs and is an essential nutrient that aids in cell growth, helps with energy generation, maintaining body temperature, protect organs, help absorb other essential nutrients that aid in producing energy, improve blood cholesterol level, help reduce inflammation in case of injury, and help in storing energy that can be used for survival when you go without food for few days [1]. But we do need to keep a track of the consumption because anything that is remotely excess leads to a variety of serious health issues [1].

### 3.1 Dietary Fats

There are different types of fat if? some are good and some are bad and some needs to be taken within a certain limit [1].

*3.1.1 Saturated Fat.* More intake of saturated fats results in the cholesterol levels in the blood which increases the risk of heart-related diseases [1]. The American Heart Association suggests around 5 percent of daily calories from foods containing saturated fat [1]. Meat, cheese, and milk are some of the sources of saturated fat [1].

*3.1.2 Trans Fat.* Any type of trans fat whether it is natural or artificial is not good [1]. The reason why food manufacturers use trans-fat is that they are less expensive, can be produced artificially, easy to use with other ingredients, last for a long time and also aid in improving the taste of the food [1]. Trans fats raise the bad fat levels and decrease the good fat levels [1]. The American Heart suggests to completely cut off trans-fat from the diet [1].

*3.1.3 Monounsaturated Fat.* Monounsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also help in gaining vitamin E which is a good nutrient that acts as antioxidant [1]. Olive oil, avocados, and sesame oil are some of the sources of monounsaturated fats [1].

*3.1.4 Polyunsaturated Fat.* Polyunsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also provide some nutrients that are essential for the body [1]. Soybean oil and sunflower oil are some of the sources of polyunsaturated fats [1].

### 3.2 Data Cleaning and Transformation

To make the analysis more interesting, the top 20 countries with most value counts for the attributes have been considered. The countries with names combined with other countries were also cleaned in the process. The data was analyzed for missing values and the attributes with more than 60 percent missing values were removed from the analysis to add consistency. Only the columns that are meaningful in the analysis were retained and the rest were removed from further analysis.

We then display the top 5 countries as a pie-chart and the 5 countries are namely United States, France, Switzerland, Germany, and Spain as shown in Figure 1.

[Figure 1 about here.]

We then impute all the null values with zeroes and we then check the dietary fat content in the foods and check the top countries with fat content using a histogram. The analysis with respect to the fat countries is as follows

### 3.3 Fat Content

The top 5 countries with most fat content in the food items are Serbia, United States, Switzerland, Germany, and Sweden as shown in Figure 2.

[Figure 2 about here.]

The top 5 countries with most saturated fat content in the food items are Serbia, United States, Germany, France and Switzerland as shown in Figure 3

[Figure 3 about here.]

The top 5 countries with most trans-fat content in the food items are United States, Brazil, Canada, Australia, Russia, and Serbia as shown in Figure 4.

[Figure 4 about here.]

The top 5 countries with most cholesterol content in the food items are United States, Canada, Portugal, Brazil, France, and Italy as shown in Figure 5.

[Figure 5 about here.]

### 3.4 Sugar and Salt Content

Although the body needs sugar, high intake of artificial and processed sugar is bad for health as it does not add any nutrients but only adds calories [5]. It is always better to rely on the natural sugar that comes with fruits and milk [5]. Artificial sugars tooth decay and diabetes [5]. Just as fat, sodium which is the main source of iodine is essential for health although its intake should be within limit [5]. Increase in intake of salt leads to blood pressure and has an effect on the heart [5].

The top 5 countries with most sugar content in the food items are United States, Serbia, Switzerland, France, and Sweden as shown in Figure 6.

[Figure 6 about here.]

The top 5 countries with most sodium content in the food items are United States, Hungary, Serbia, Sweden, and France as shown in Figure 7.

[Figure 7 about here.]

## 4 NUTRITION GRADE LABELLING SYSTEM

France recently took a decision to implement a nutri-score system which will use a color coding mechanism to label the food products that will help consumers know the nutrition grade of the product [3]. The World Health Organization regional office for Europe as a part of its 5-year action plan from 2015-2020 recommends a labeling mechanism for the consumers to know about the quality of the food products at a first glance [3]. This will not only make it easier for the consumers to pick healthier options but it will also regulate food manufacturers to resort to healthier ingredients instead of going for low cost artificial or less healthy ingredients [3].

France after the United Kingdom became the second country to implement this system to indicate the main ingredients like fat, salt and sugar content in the food items [3]. France made use of an evidence-based system to study different labeling systems to arrive at the best one [3]. By implementing this system, the World Health Organization will keep a check on the growing number of diet-related diseases in the Europe region [3]. Europe being the largest consumer of cheese wants to regulate the ingredients that go into the manufacturing process so that people are well informed about their food choices [3].

### 4.1 Nutrition Grade Prediction as a Big Data Problem

We build a predictive classification model to predict the food nutrition grade based on the ingredients of the food. The goal is to apply various machine learning algorithms to the problem at hand, measure the prediction accuracy to compare and contrast the different algorithms and arrive at the best algorithm that suits the given data and the problem. This problem can be solved using Big Data and Machine Learning techniques given the size and the complexity of the data.

## 5 MACHINE LEARNING

Machine Learning is a field in which we train computers in a way that they can learn from the input data [6]. The ideology is that computers use the training data that is made available to them, learn from it, build a model and use this experience to build knowledge that can be applied on new unseen data [6]. A wonderful example to demonstrate machine learning is the application to detect spam emails where the machine builds knowledge from previously seen emails which are marked as spam, checks new emails to see if they match the historic spam emails and label them as spam or non-spam [6].

## 5.1 Types of Machine Learning Algorithms

There are primarily two types of machine learning algorithms, descriptive models and predictive models [6]. A *Descriptive Model* is described as the analysis done and insights gained from slicing and dicing the data in new and interesting ways [6]. One example of a descriptive model is pattern discovery that is often used in market basket analysis where transnational purchase details are analyzed [6]. A *Predictive Model* on the other hand involves predicting one value using one or more variables [6]. The learning algorithms tried to build a model that captures the relationship between a response variable and the independent variables [8].

## 5.2 Types of Learning

*Unsupervised Learning* is the process where there is no explicit training data to learn from, so there is simply no mechanism where the machine can learn from previously available data [6]. The same email example can be looked at in a different way where we now want to do anomaly detection in emails [6]. Here the main goal is to detect unusual messages from the bunch of messages and we do not have experience of previous data [6].

*Supervised Learning* in contrast is the process of gaining knowledge or expertise from the training data which can be applied to future unseen data [6]. Here the model is first trained by using a bulk of training examples and this model is applied to testing data to measure the accuracy [6]. The variable that we need to predict is identified which is called the response variable and the variables that are used to predict the response variables, called the predictor variables are identified [6]. If the existing variables are not sufficiently giving the accuracy that is expected, a method called feature engineering is done where new variables are derived by combining existing variables [6].

## 6 PREDICTION ANALYSIS

Prediction analysis is the process of working on a large dataset using a combination of statistical, data mining and machine learning algorithms to predict the outcome based on past data [6]. There are primarily two types of prediction analysis in machine learning, namely regression and classification [8]. In regression, we try to predict a continuous variable from the predictor variables [8]. A good example of regression is to predict the housing prices from different parameters like the year of construction, location, amenities, number of bedrooms etc [8]. Here the response variable is continuous and it is not predefined [8]. Classification, on the other hand, tries to predict a categorical variable in which we assign each record with a predefined label or a class [8].

Classification is the task of assigning each data record to a predefined class [8]. In machine learning, classification is categorized as a supervised learning technique [8]. This problem has applications in various fields like spam detection, medical applications, astronomy, and banking to identify fraudulent transactions from genuine transactions [8]. It is the task of coming up with a model which is essentially a function that maps every data record to a class label [8].

The task at hand is a classification problem since we are trying to predict the food nutrition grade of the products based on the ingredients that go into the product. For this problem, we are considering only the data for the country France, since the nutrition grade is available for most food products from the country. Another reason is that France is the first country in the region to come up with the idea of adding a color-coded label to the food products mentioning the nutrition grade. In the subsequent sections, we discuss the machine learning techniques used to solve this problem.

## 6.1 K Nearest Neighbors

**6.1.1 Overview.** Some of the classification algorithms in machine learning work on the principle of eager learning that involves a two-step process where first a model is built from the training data and the model is applied on testing data [8]. In contrast, K nearest neighbors is a lazy learning algorithm where the process of modeling the training data is not done until the test examples are classified [8]. *Rote Classifier* is a good example of lazy learning algorithm which memorizes the entire training data to perform classification but has the drawback of not being able to map every test example against the training example [8]. K nearest neighbors algorithm overcomes this drawback by finding all the records that are closest or nearest to the training records [8].

The nearest neighbor puts each attribute list as a data point in the n-dimensional space, given n the number of attributes [8]. Once we have the training examples, we take each test example and compute its distance to the training example classes and assign a class label [8]. Any of the popular distance measures among Euclidean distance, Manhattan distance, Minkowski distance and Mahalanobis distance can be used [8]. The k denotes the k closest points to the test example [8]. Figure 8 shows the algorithm [8].

[Figure 8 about here.]

**6.1.2 Support in Python.** KNeighborsClassifier is available in the scikit learn python library.

## 6.2 Logistic Regression

Logistic regression or logit regression is a special type of regression analysis where the response variable that we need to predict is a categorical variable [8]. Typically, logistic regression models the response variable to take two values, 1 or 0, pass or fail, win or lose [8]. Logistic regression that takes more than two values for the response variable is called multinomial logistic regression [8]. Here the probability of the response variable to take a categorical value is modeled as a function of the predictor variables [8].

Like a lot of machine learning algorithms, logistic regression works by making a lot of assumptions which should be taken care as a part of the data cleaning and transformation process [6]. It does not assume a linear relationship between the response variables and the predictor variables [6]. Since it applies a log transformation on the predicted probabilities, it can handle a variety of relationship between the predictor variables [6]. If the predictor variables are multivariate normal, the algorithm achieves the best result although it works even if they are not [6]. The stepwise method must be used in the logistic regression to ensure that we are neither overfitting

nor underfitting the data [6]. A very important assumption to be noted in logistic regression is that each attribute list must be independent, in the sense, the data records must not be derived from a before-after setup experiment [6]. It also requires a decently large sample size to work on [6].

**6.2.1 Support for Python.** LogisticRegression is available in the scikit learn python library.

### 6.3 Random Forest Classifier

Random forest is an ensemble classification algorithm which is very powerful [8]. Ensemble method is a special process to improve the accuracy of the prediction [8]. The classification algorithms we have seen so far predict the response variable using a single classifier on the test data but ensemble methods use multiple classifiers in tandem and aggregate the predictions to boost the accuracy by a huge margin [8]. Using a combination method, the ensemble method derives a set of base classifiers from the training data and on each iteration takes a vote of all the base classifiers to arrive at a result [8].

Random forest is an ensemble method which works very well for classification problems [8]. It combines the predictions made by multiple classifiers where each classifier independently works on the training data and casts its vote [8]. Unlike methods like AdaBoost which generates values based on independent random vectors using a varied probability distribution, random forest generates values based on fixed probability distribution [8].

**6.3.1 Rationale for Random Forest.** Consider an example, where we have 25 base classifiers and each base classifier has an error rate of 0.35 [8]. As discussed, the random forest takes the majority vote given by the base classifiers [8]. The model makes a wrong prediction if half or more base classifiers predict inaccurately. The accuracy is improved with an error rate of 0.06 which is far better than using just a single classifier [8].

**6.3.2 Support for Python.** RandomForestClassifier is available in the scikit learn python library.

## 7 EXPERIMENTS AND RESULTS

In this section, we will introduce the algorithm along with the details of experiments and methodology for predicting the nutrition grade of food products in France.

### 7.1 Algorithm

The problem at hand is to correctly identify the nutrition grade of the food item. The possible labels are, *a* to *e*, with *a* being the best and *e* being the worst grade for a food item. For this task, we have used machine learning techniques that help in predicting the label of each food item. Before getting into the details of each step of the method, we first present a concise version of the algorithm used for this task:

- (1) Select all the records for the country, France. Drop records where nutrition grade is not populated.
- (2) Separate the predictors from the response variable in order to perform data cleaning and data transformation steps.

- (3) Check for missing values in the predictors obtained in the step above. Drop columns with more than 60% missing values.
- (4) Impute the missing values with 0 for remaining columns.
- (5) After imputing the missing values, standardize all the numerical predictors using the standard scaler.
- (6) Check for the correlation between different numerical predictors. Drop one predictor from each pair of predictors that show high correlation.
- (7) Combine the pre-processed predictors and the response variable in a single data frame.
- (8) Divide the data obtained in step above into training and test data using stratified sampling.
- (9) Train different classifiers on the training data and check the performance of each classifier on the test data.

### 7.2 Data set

For the classification problem, we selected the records for country France.

Number of examples: 123,961

Number of variables: 12

Response variables: *Nutrition Grade*

Predictor variables: *Energy per 100g*, *Fat per 100g*, *Saturated Fat per 100g*, *Carbohydrates per 100g*, *Sugars per 100g*, *Fiber per 100g*, *Proteins per 100g*, *Salt per 100g*, *Trans-fat per 100g*, *Sodium per 100g*

### 7.3 Python Packages Used

The following Python packages were used to solve the classification problem:

- Pandas: Provides high-performance data structures for data analysis and data munging
- Matplotlib: Plotting library that helps to embed plots into applications using GUI
- Seaborn: Visualization package based on matplotlib used for drawing high-level statistical graphics
- Scikit-learn: Toolbox with solid implementation of machine learning and other algorithms
- Scipy: Package that supports scientific computing with modules for linear algebra and integration

### 7.4 Data Cleaning

**7.4.1 Step 1: Data Sparsity.** Data sparsity refers to the situation where a lot of attributes have missing values which is an advantage in some cases because you only need to store and analyze the data that is available to you and save on computation time and storage [8]. We first check the data value counts for each country. United States, France, Switzerland, Germany, and Spain come as the top 5 countries with most data. Since the food nutrition grade was implemented in France, it has most products for which nutrition grade is labeled. So for this classification problem, we use the food data from France for analysis.

**7.4.2 Step 2: Handling Missing Values.** Missing values is a common scenario and they can be handled in different ways. You could

choose to eliminate the data objects with missing values but at the expense of missing some critical analysis [8]. Estimating the missing values is also a good way to handle them, especially when the data comes from time series etc, where you could possibly interpolate the missing values from the ones that are closer to it [8]. Ignoring the missing values is another technique which can be applied to tasks like clustering where the similarity can be calculated using the attributes other than the missing ones [8].

The data set was first analyzed to check the missing values in all the columns. The threshold limit has been set at 60 percent. All the columns with missing values more than 60 percent were removed from the analysis to make the result more consistent. Once the columns were removed, the data set has to be re-indexed to maintain the order. Only the columns that are important for the prediction task have been retained from the original dataset. In this case, all the ingredients which are primarily the predictor variables were included. The missing values in the response variable also need to be taken care of. Removing the records with missing values for the response variable proved to be the best option for trying out various things.

Imputation was used to handle the null values in the predictor variables. Imputation can be done in a variety of ways, for example, replacing the missing values with zero or imputing the missing values for numerical columns with the mean and the categorical columns with the mode. Since all the predictor variables have numeric values, all the null values have been replaced with zero. To ensure that the imputation process has been done correctly, the sum of missing values is calculated since post-imputation, this sum should be zero.

**7.4.3 Step 3: Outlier Treatment.** Outliers are data objects with quite distinct characteristics from the other data records [8]. There is a considerable difference between anomalies and outliers, where anomalies refer to data records that have bad data, which is noise and need to be ignored, anomalies often contain interesting aspects and can lead to some good analysis [8]. In applications like *Fraud Detection*, anomalies could be of utmost importance [8]. The outliers in the data have been looked at by using box plots and have been handled as a part of the data cleaning process.

## 7.5 Exploratory Data Analysis

For exploratory data analysis, we used the Seaborn package along with Matplotlib for visualizations. The measure of spread, that is the range and variance of the values, is a good way to understand the different aspects of the predictor variables. Box-plots are a method of visualization to look at the distribution of values for a numerical attribute [8]. The box plots show the percentiles where the lower and upper ends of the box indicate 25<sup>th</sup> and 75<sup>th</sup> percentile, the line inside the box indicates the 50<sup>th</sup> percentile, the tails indicate the 10<sup>th</sup> and 90<sup>th</sup> percentile respectively [8].

**7.5.1 Bi-variate box-plots.** Bi-variate box-plots go beyond univariate box plots by showing the relationship between the predictor variable and the response variable [8]. We look at the bi-variate

box-plots for each of the important predictor variables namely, saturated fat, polyunsaturated fat, sugars and salt and the response variable, nutrition grade. Figure 9 shows the bi-variate box plots.

[Figure 9 about here.]

By looking at the box plots, we can understand some important aspects of how the response variable is related to the predictor variables. We see that as the average saturated fat content increases, the food grade decreases and as the average polyunsaturated fat content increases the nutrition grade is better. When the sugar levels increase, the health quotient of the food comes down. The energy levels behave in an interesting manner where the energy for the nutrition grade A is higher whereas in general, the average energy level slightly increases with the decreasing nutrition grade. While increase in energy does not necessarily imply that the nutrition quality is high, as there are a lot of instant energy foods that have a lot of additives, but they are often rated low when it comes to health.

**7.5.2 Correlation.** Correlation between data objects is the measure of the linear relationship between the attributes of the object that are continuous variables [8]. Correlation analysis is the process of finding of the correlations between the different predictor variables and identify high collinearity problem [6]. The relationship could be either linear or non-linear based on the given data [8]. The correlation coefficient can range anywhere between -1 and 1, where 1 indicates a very high positive correlation and -1 indicates a very high negative correlation [6]. Correlation plot visually shows the correlation coefficient between the variables in a nicely laid out plot. Figure 10 shows the correlation plot.

[Figure 10 about here.]

By looking at the correlation plot, we can see that sugars, fat, energy are positively correlated with the nutrition grade. This indicates that these variables will play an important role in the prediction algorithm. However, sodium and salt are highly correlated with each other and this may lead to collinearity problem if not handled. Collinearity is the state where the independent variables are highly correlated with each other which can add a lot of noise to the data [7]. Some of the problems because of collinearity are that the regression coefficients may not be estimated correctly. Also, collinearity makes it very difficult to explain the response variables using the predictor variables [7]. So we remove sodium from the predictor variables and proceed to the next step.

**7.5.3 Data Transformation.** Data transformation refers to the transformation that is applied to the variables [8]. For each data object, we apply a transformation function to all the attributes of the object to ensure that the attributes do not have a lot of variance in the data [8]. This process is also called standardization since we are applying a standard function to make sure all the attributes fall within a given range [8]. There are different methods that can be applied to achieve scaling namely log transformation, absolute value, square root transformation [8].

We use the method called normalization where all the values fall in the range, 0 to 1. To achieve this, we use the prepossessing package from sklearn which provides utility functions and transformer classes to change raw data into a standard representation. A lot of machine learning algorithms work well on standardized data. If some of the variables have extreme values, they might dominate the model function and might disturb the estimation parameter. Thus, for such extreme values, standardization helps achieve better results.

On scaling the data, there was a massive improvement in the prediction accuracy of the algorithms, implemented for this task. Thus, this proves the importance of data standardization with respect to machine learning algorithms.

## 7.6 Data Sampling

In a supervised machine learning approach, the model is trained on one sample of the data and later tested on a different sample of the data. Thus, in order to test the performance of the nutrition grade classifier, the data for the country France was divided into two samples, training and testing. There are various ways to achieve this split or sampling of the data. Some of these sampling methods are:

- Simple Random Sampling: This is one of the simplest sampling techniques. In this technique, every data point has an equal chance of being selected. In other words, it works similar to a lottery system where every outcome has an equal probability. The biggest advantage of this technique is the ease of implementation and its unbiased nature while generating the sample. However, random sampling might not always result in a sample that can represent the true population. It generally works well when we have huge data to sample from.
- Stratified Sampling: This technique is a more sophisticated method of sampling data. Stratified sampling generates a sample such that the proportion of each class in the sample is same as that in the true population. In this technique, the entire population is divided into groups or strata. The next step is to randomly select data points from each stratum such that the final sample has the same proportion for each stratum as that present in the true population. Thus, the sample generated by this technique is a good representative of the true population. Stratified sampling is a very useful technique when the classes in the data are highly imbalanced.

For our classifier, we chose to divide the data for France into training and test samples using stratified sampling technique. The strata or groups were created based on the response variable, i.e., food grade. This ensured that the training and test data had the same proportion of each food grade.

## 7.7 Data Modeling

Once the data was divided into training and test data, the next step was to train different classifiers and tune their respective parameters for better accuracy. We implemented three different models for

classifying the food grade. Each of these models along with their parameters is:

- K Nearest Neighbors (kNN): For kNN, the grade of a food item in test data is classified by first finding the  $k$  most similar food items in the training data. It then takes the vote (food grade label) from each of these neighbors and based on the majority vote, the food item from the test data is assigned a food grade. Thus, one of the most important parameter for kNN is  $k$ , i.e., the number of neighbors to consider from the training data. We tried different  $k$  values and found that  $k = 3$  gives the best accuracy.
- Logistic Regression: For logistic regression, one of the important parameters is the penalty. This parameter specifies the kind of regularization to be applied. This parameter can take two possible values,  $l_1$  regularization and  $l_2$  regularization. Both these values penalize high magnitude of the coefficients of the predictors in order to prevent the model from over-fitting. For our model, we have used  $l_2$  regularization as it works well even in the presence of highly correlated features.
- Random Forest: For the random forest, there are many parameters, such as the number of trees in the forest, the maximum depth of the trees, maximum number of features to consider at each split, the minimum number of samples required in a sub-tree to qualify for a further split, the minimum number of samples required to qualify as a leaf node, etc. For our data, we have kept most of the parameters at their default values, except for, the number of estimators or trees in the forest. We have set this value to 100, as the classifier resulted in very high accuracy with 100 trees in the forest.

## 7.8 Evaluation Metrics and Results

There are various evaluation metrics for assessing the performance of classifiers. Some of these evaluation metrics are [4]:

- Accuracy: This metric gives the proportion of the total number of correctly classified instances
- Precision: This gives the proportion of the true positive instances from the total instances classified as positive
- Recall: This gives the proportion of the positive instances that are correctly classified
- F-Measure: This gives the harmonic mean between precision and the recall values
- Confusion Matrix: This is a useful way of checking the accuracy of the classifier. It clearly shows the number of instances correctly classified for each label. Thus, if we know that the classes in the data are not well-balanced, it's always a good idea to check the confusion matrix along with accuracy. Consider a case where 95% of the instances belong to class A and only 5% of the instances belong to class B. If a classifier is trained on a dataset with such imbalance, there is a high chance that the classifier would return label A for each test instance. The classifier would still be able to correctly classify 95% of the test instances resulting in 95% accuracy. This is a case where accuracy can be misleading and thus a quick look at the confusion

matrix can help understand the problem with the classifier. For such a case, the confusion matrix will clearly show that all the instances of the minority class, B, have been misclassified.

For our model, we used accuracy as well as confusion matrix for evaluating the results. The confusion matrix did not show any serious issues for any of the classifiers. The accuracy for each of the three classifiers was:

- (1) Logistic Regression: With  $l_2$  penalty, the accuracy of logistic regression was 78.9%. Figure 11 shows the confusion matrix.

[Figure 11 about here.]

- (2) K Nearest Neighbors: With k as 3, the accuracy of kNN was 95.74%. Figure 12 shows the confusion matrix.

[Figure 12 about here.]

- (3) Random Forest: With a number of trees as 100, the accuracy of random forest classifier was 99.68%. Figure 13 shows the confusion matrix.

[Figure 13 about here.]

Thus, we obtained the best results with Random Forest classifier.

## 8 CONCLUSION

Analysis of food content is very important in today's world as most of the companies try to fool consumers by labeling their product as low-fat. It's important for the consumers to know the true nutrition grade while purchasing any food item. Thus, we analyzed the nutrition grade based on the composition of various components of the food items. We developed a model that labels a food item purely on the basis of its nutrients, thus eliminating any bias, such as, the production company or the brand name. For accurate labeling, we applied different data cleaning and data transformation techniques. With this transformed data, we tried various machine learning models. We got the best results using random forest classifier which was able to accurately label 99% of the food products. Since the model is trained only for France, as part of future work, we can try and scale our model for different countries. However, to achieve similar results for other countries, we need to collect more data. The current data has many missing values for countries other than France. Once we collect enough data for these countries, we can also try and implement more sophisticated models like neural networks in future.

## ACKNOWLEDGMENTS

This project was undertaken as a part of the course objective for I523: Big Data Applications and Analytics at Indiana University, Bloomington. We would like to thank Dr. Gregor von Laszewski and all the TAs for their help, support, and suggestions.

## A WORK BREAKDOWN

**Dataset identification:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Requirement Gathering:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Learning Machine Learning Concepts:** Karthik Vegi, Nisha Chandwani: work equally split between.

**Data analysis and implementation of the Logistic Regression:** Karthik Vegi.

**K nearest neighbors and Random Forest algorithms:** Nisha Chandwani

**Writing the project report:** Karthik Vegi, Nisha Chandwani: work equally split between.

## REFERENCES

- [1] American Heart Association. 2017. Dietary Fats. Webpage. (March 2017). <https://healthyforgood.heart.org/eat-smart/articles/dietary-fats>
- [2] Alejandro Cifuentes. 2012. Food analysis: present, future, and foodomics. *ISRN Analytical Chemistry* 2012 (2012), 16.
- [3] World Health Organization Europe. 2017. Labelling systems to guide consumers to healthier options. Webpage. (March 2017). <http://www.euro.who.int/en/countries/france/news/news/2017/03/france-becomes-one-of-the-first-countries-in-region-to-recommend-colour-coded-front-of-pack-nutrition-labelling-system>
- [4] M Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.
- [5] Healthy Eating SFGate. 2017. Recommended Daily Allowances of Fats, Sugars, Sodium for Adults. Webpage. (2017). <http://healthyeating.sfgate.com/recommended-daily-allowances-fats-sugars-sodium-adults-2976.html>
- [6] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, USA.
- [7] Statistics Solutions. 2017. Multicollinearity. Webpage. (March 2017). <http://www.statisticssolutions.com/multicollinearity/>
- [8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Pearson, Boston, USA.
- [9] Karthik Vegi and Nisha Chandwani. 2017. Code base - Analysis on food products around the world. github. (Dec. 2017). <https://github.com/bigdata-i523/hid231/tree/master/project/code>

#### LIST OF FIGURES

1	Top 5 countries [9]	9
2	Top 5 countries with most fat content [9]	10
3	Top 5 countries with most saturated fat content [9]	11
4	Top 5 countries with most trans-fat content [9]	12
5	Top 5 countries with most cholesterol content [9]	13
6	Top 5 countries with most sugar content [9]	14
7	Top 5 countries with most sugar content [9]	15
8	K nearest neighbors algorithm[8]	15
9	Bi-variate box plots [9]	16
10	Correlation Plot [9]	17
11	Confusion matrix for Logistic Regression [9]	18
12	Confusion matrix for K Nearest Neighbors [9]	19
13	Confusion matrix for Random Forest [9]	20

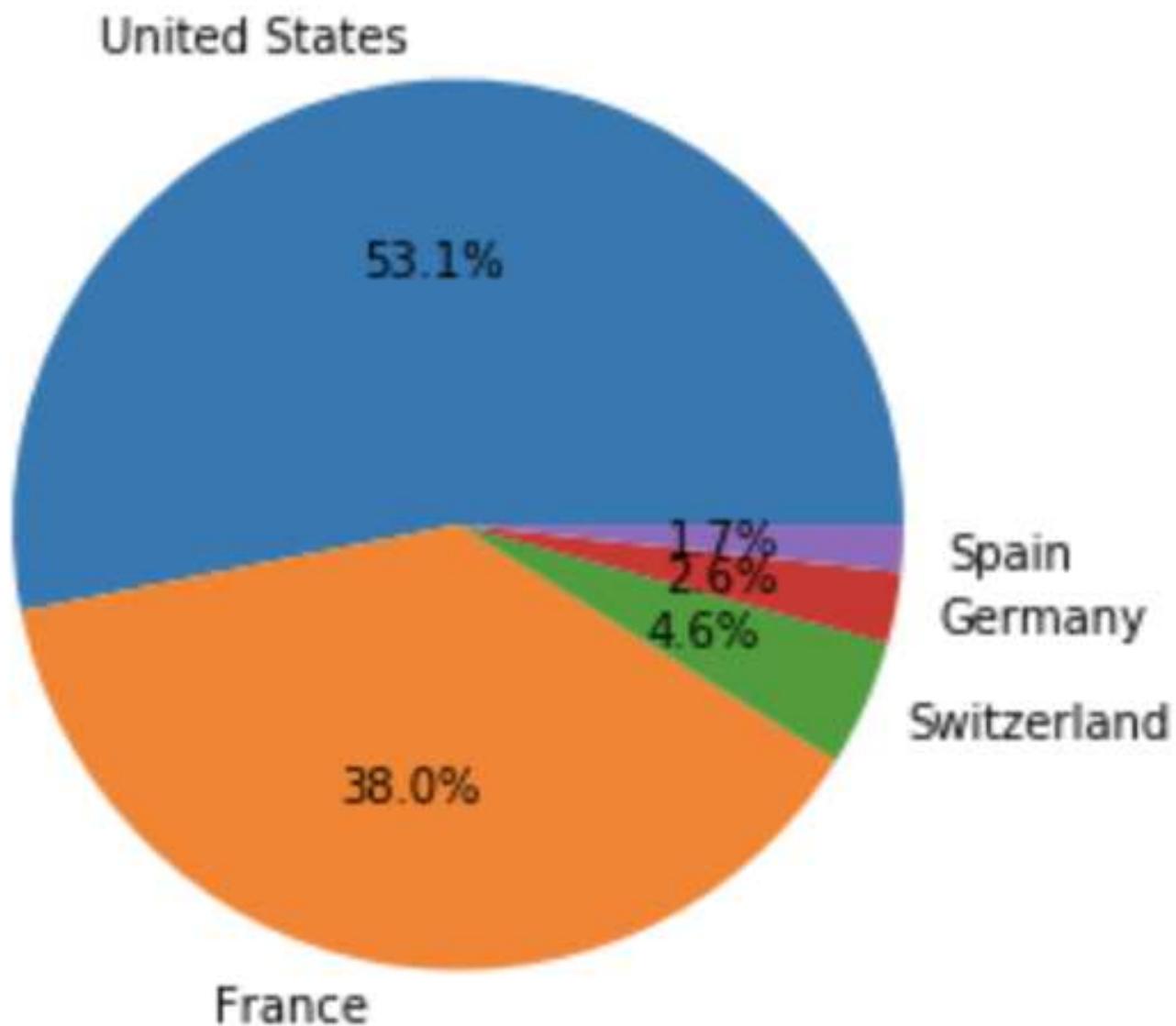


Figure 1: Top 5 countries [9]

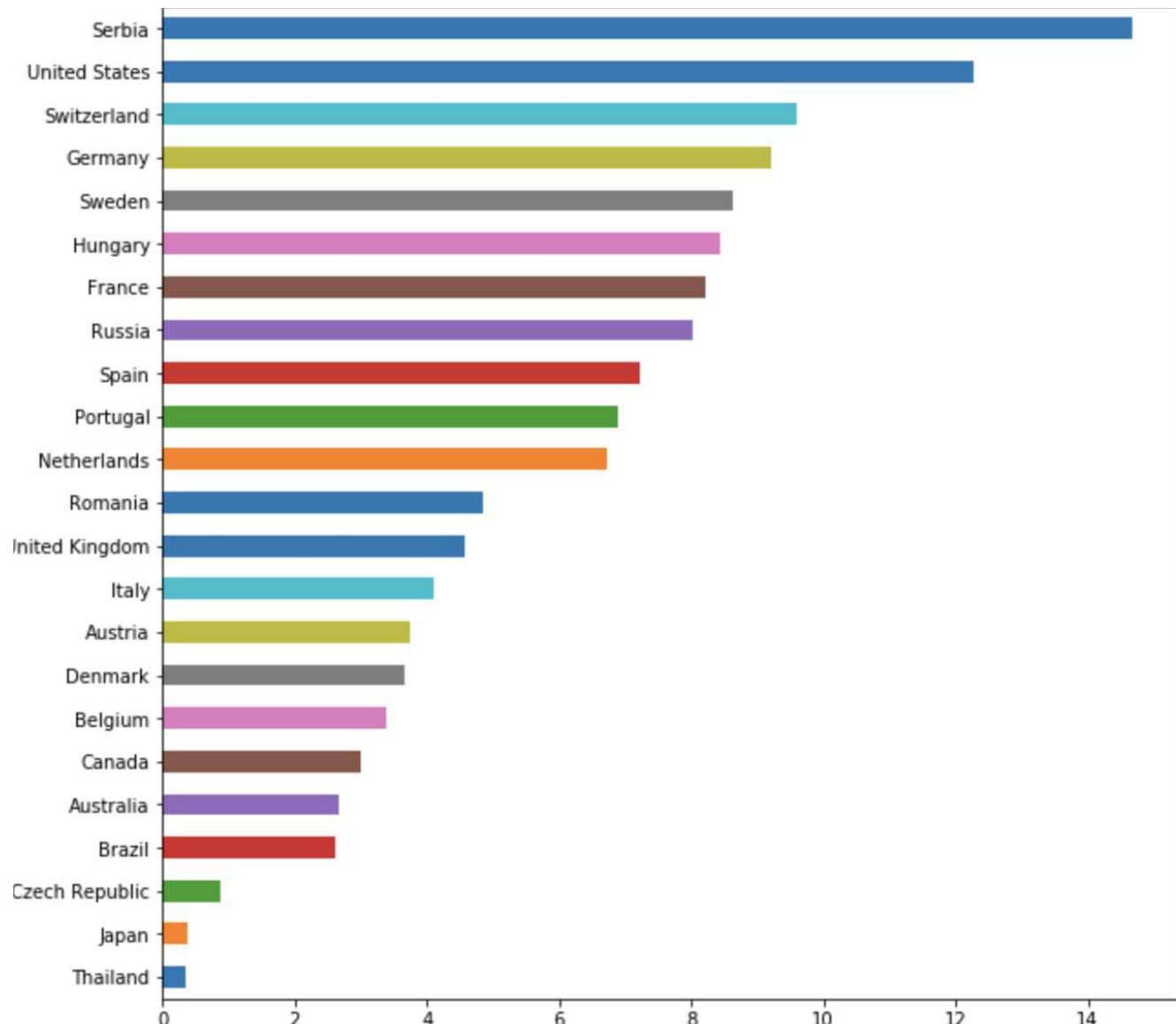


Figure 2: Top 5 countries with most fat content [9]

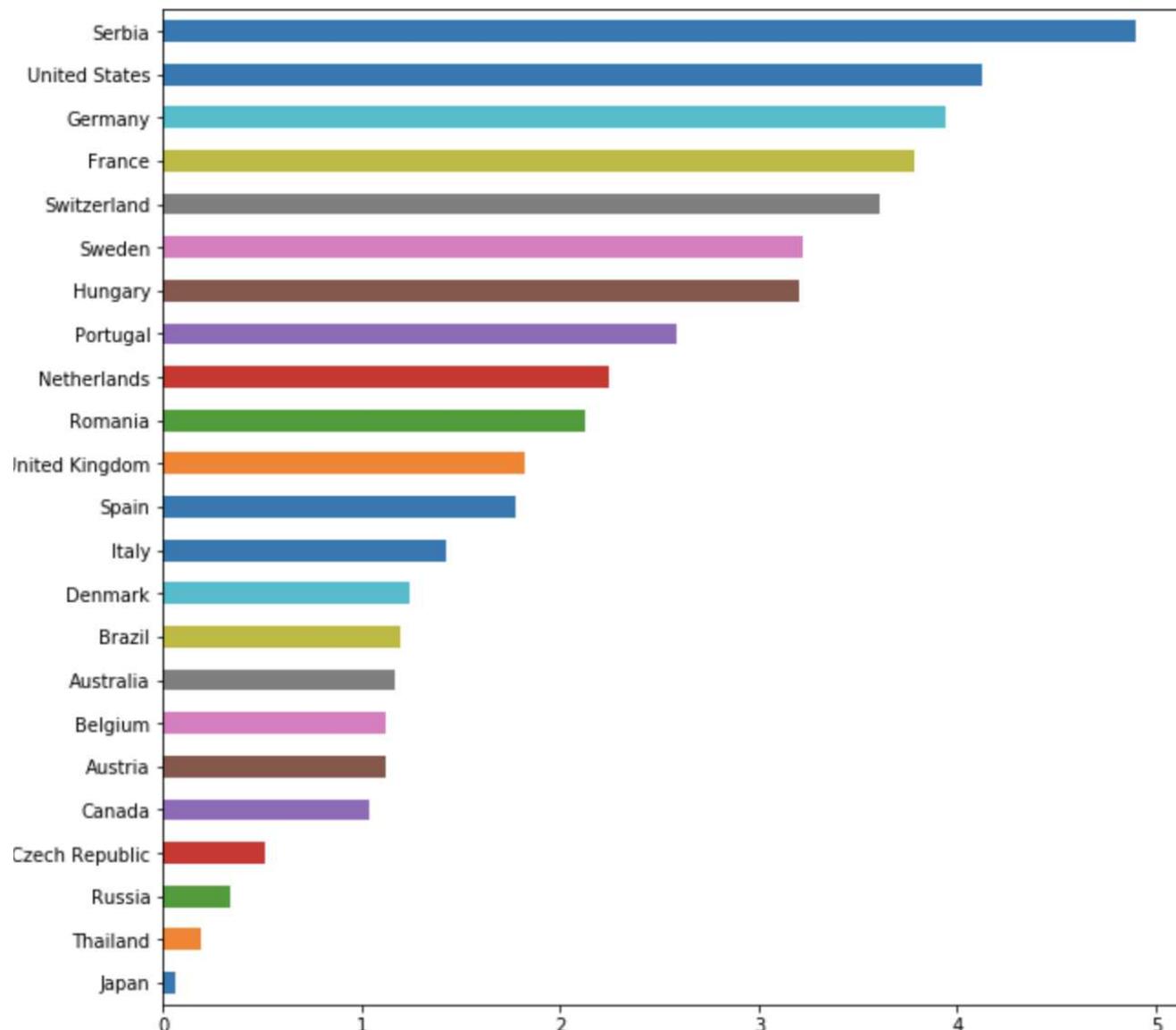


Figure 3: Top 5 countries with most saturated fat content [9]

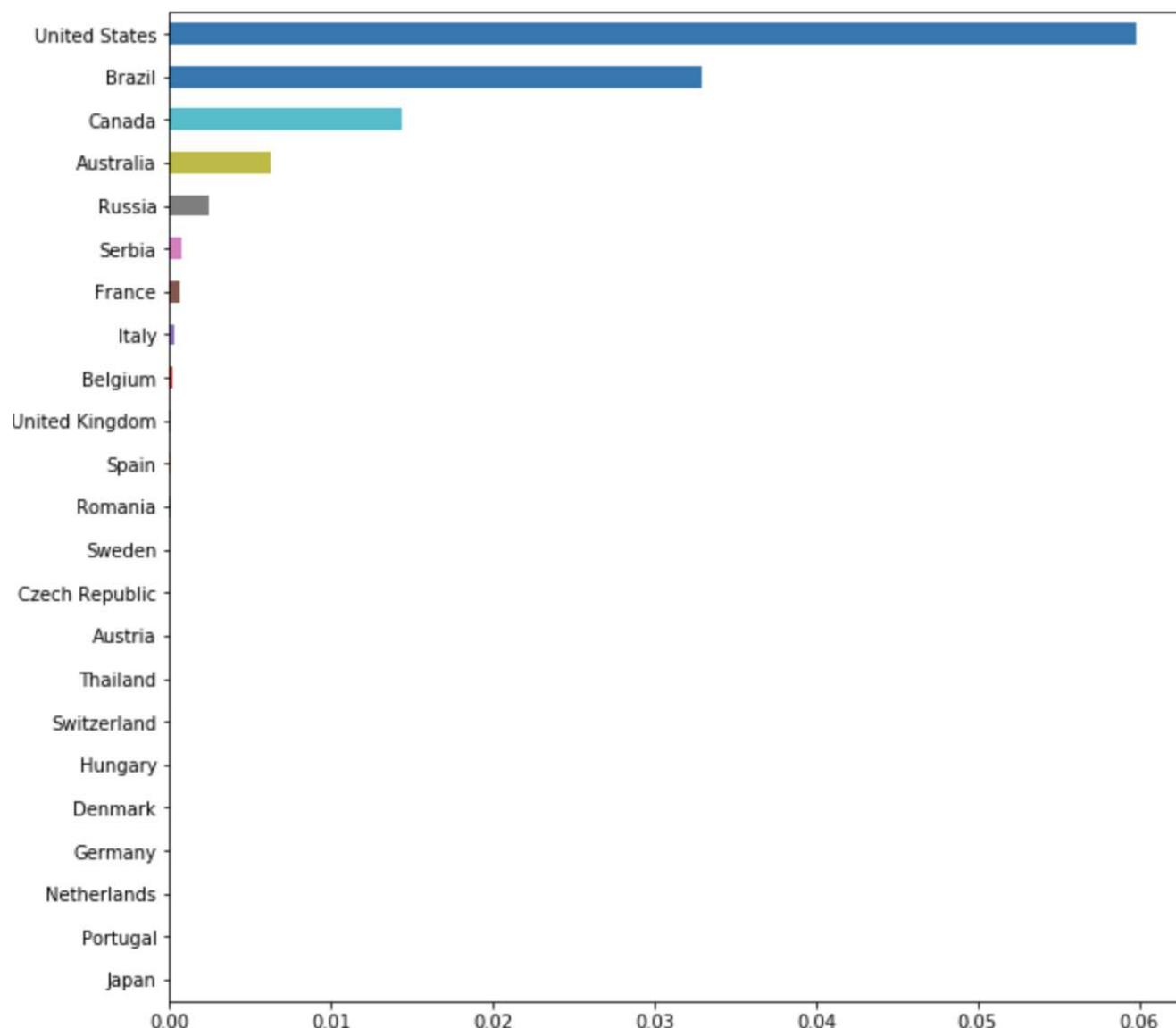


Figure 4: Top 5 countries with most trans-fat content [9]

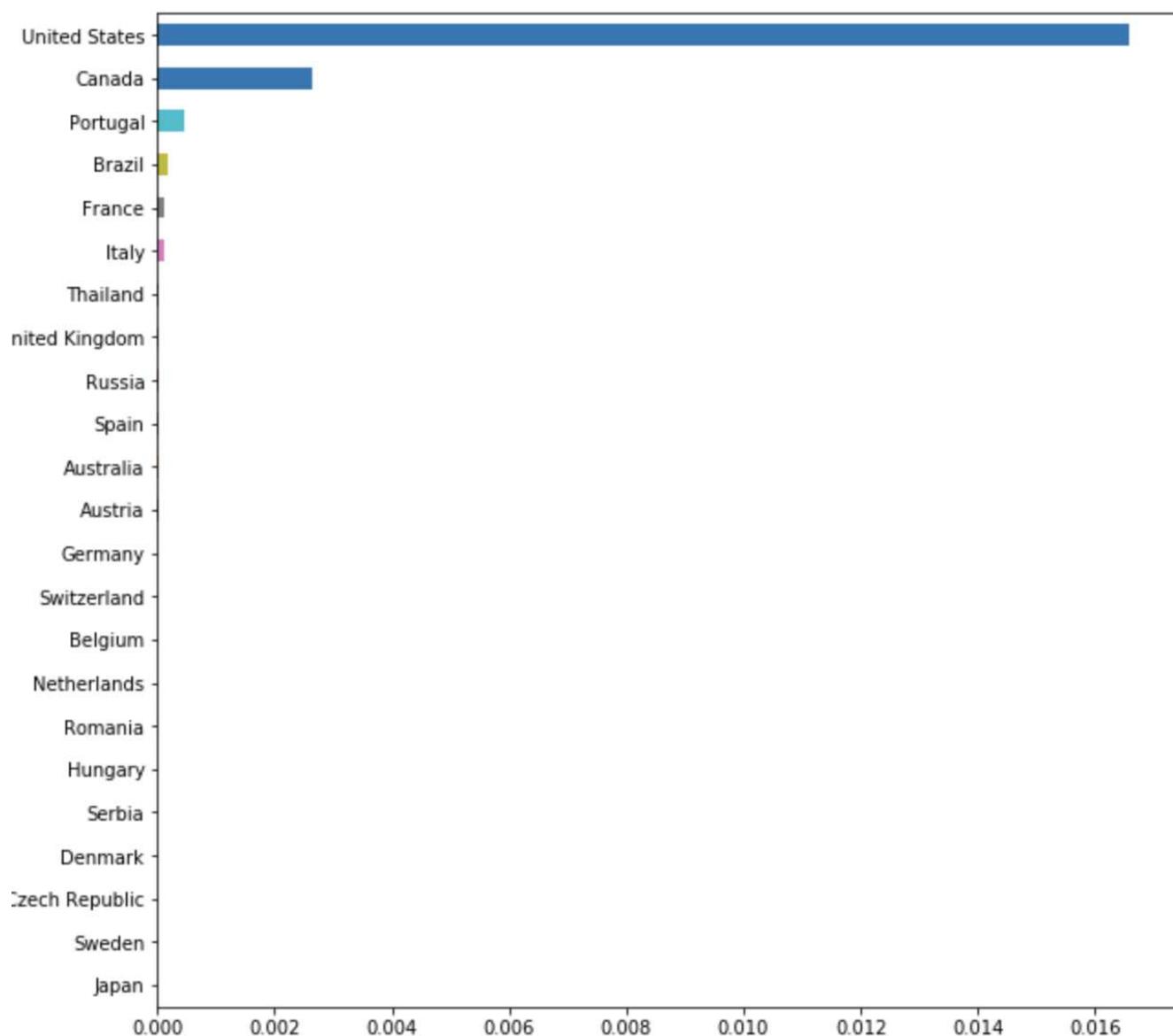


Figure 5: Top 5 countries with most cholesterol content [9]

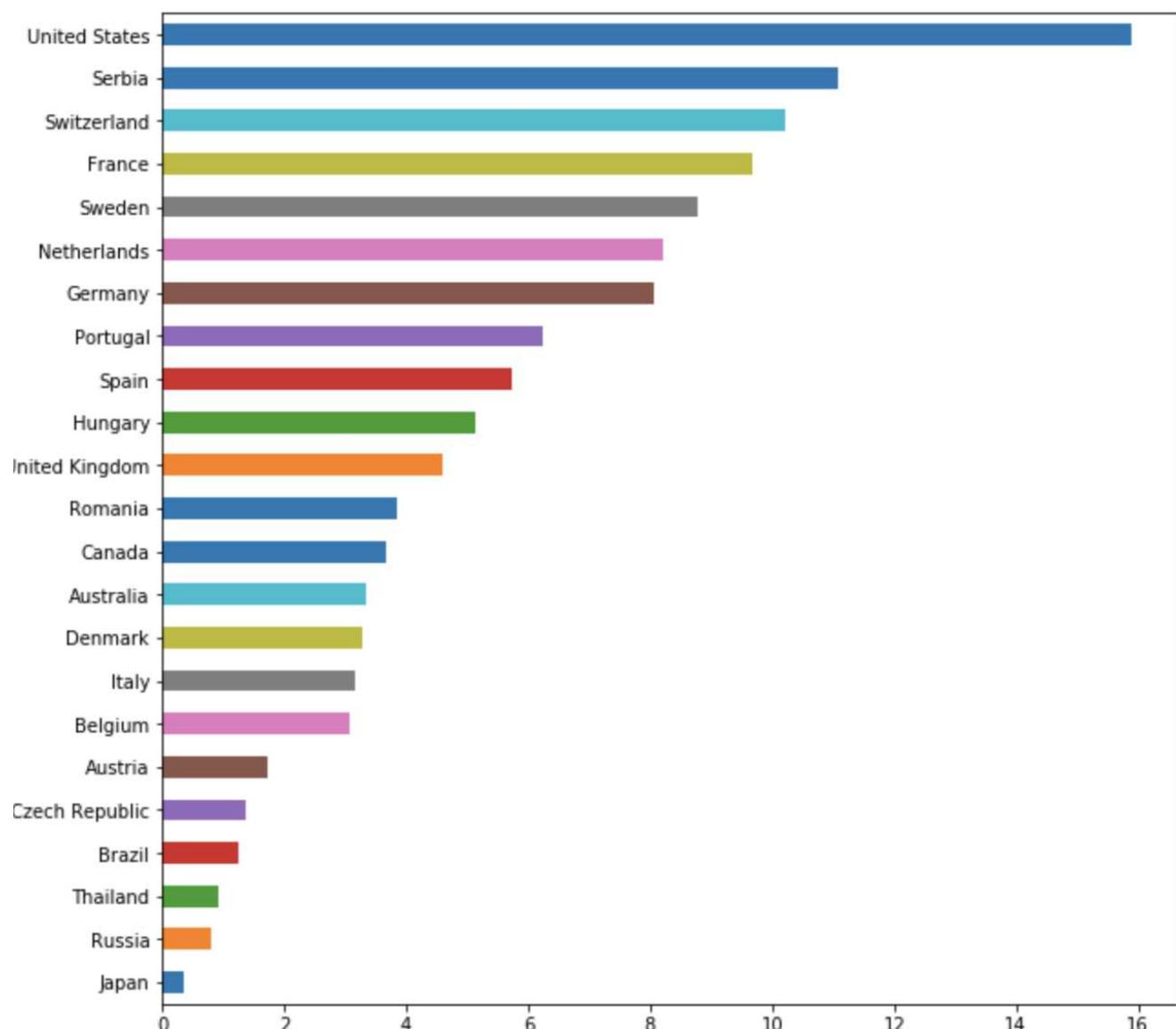


Figure 6: Top 5 countries with most sugar content [9]

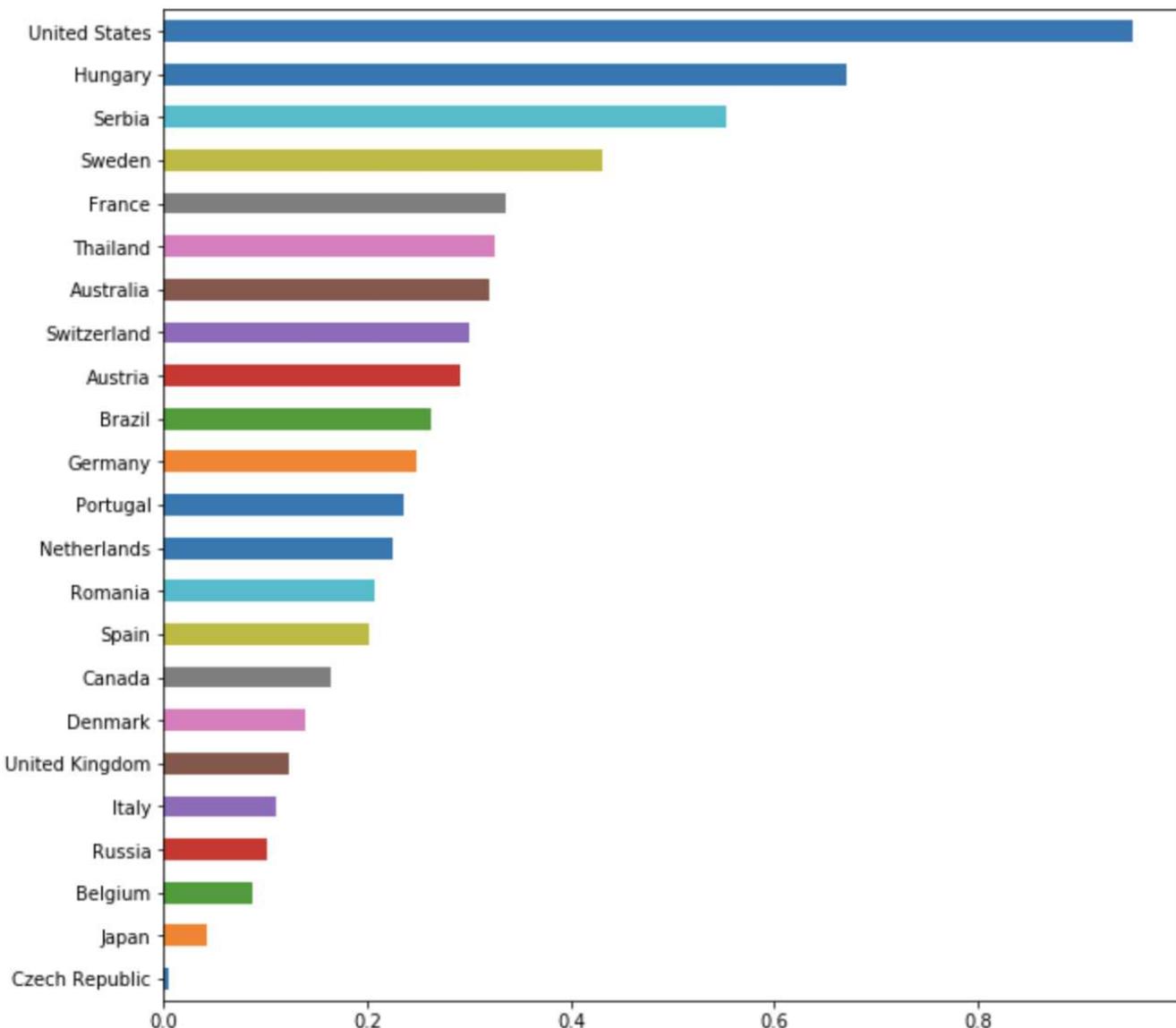


Figure 7: Top 5 countries with most sugar content [9]

---

### Algorithm 5.2 The $k$ -nearest neighbor classification algorithm.

---

- 1: Let  $k$  be the number of nearest neighbors and  $D$  be the set of training examples.
  - 2: **for** each test example  $z = (\mathbf{x}', y')$  **do**
  - 3:   Compute  $d(\mathbf{x}', \mathbf{x})$ , the distance between  $z$  and every example,  $(\mathbf{x}, y) \in D$ .
  - 4:   Select  $D_z \subseteq D$ , the set of  $k$  closest training examples to  $z$ .
  - 5:    $y' = \operatorname{argmax}_v \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
  - 6: **end for**
- 

Figure 8: K nearest neighbors algorithm[8]

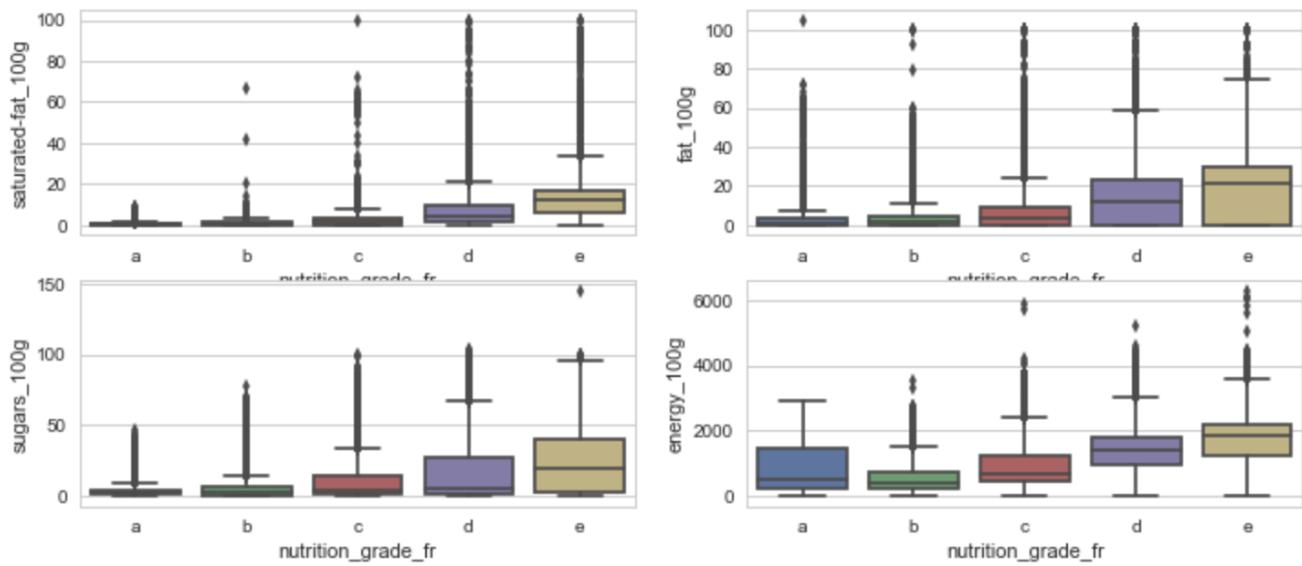


Figure 9: Bi-variate box plots [9]

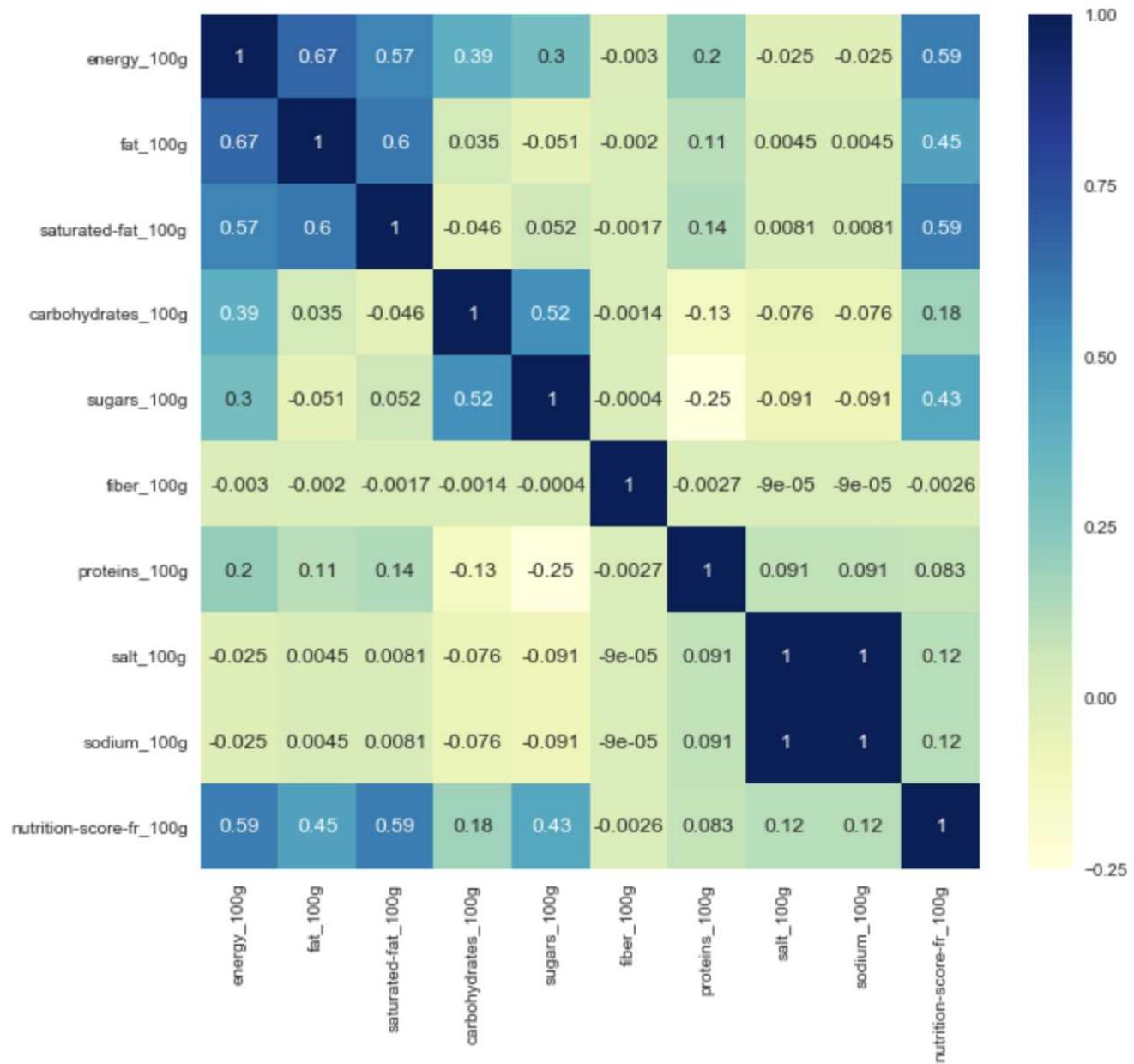


Figure 10: Correlation Plot [9]

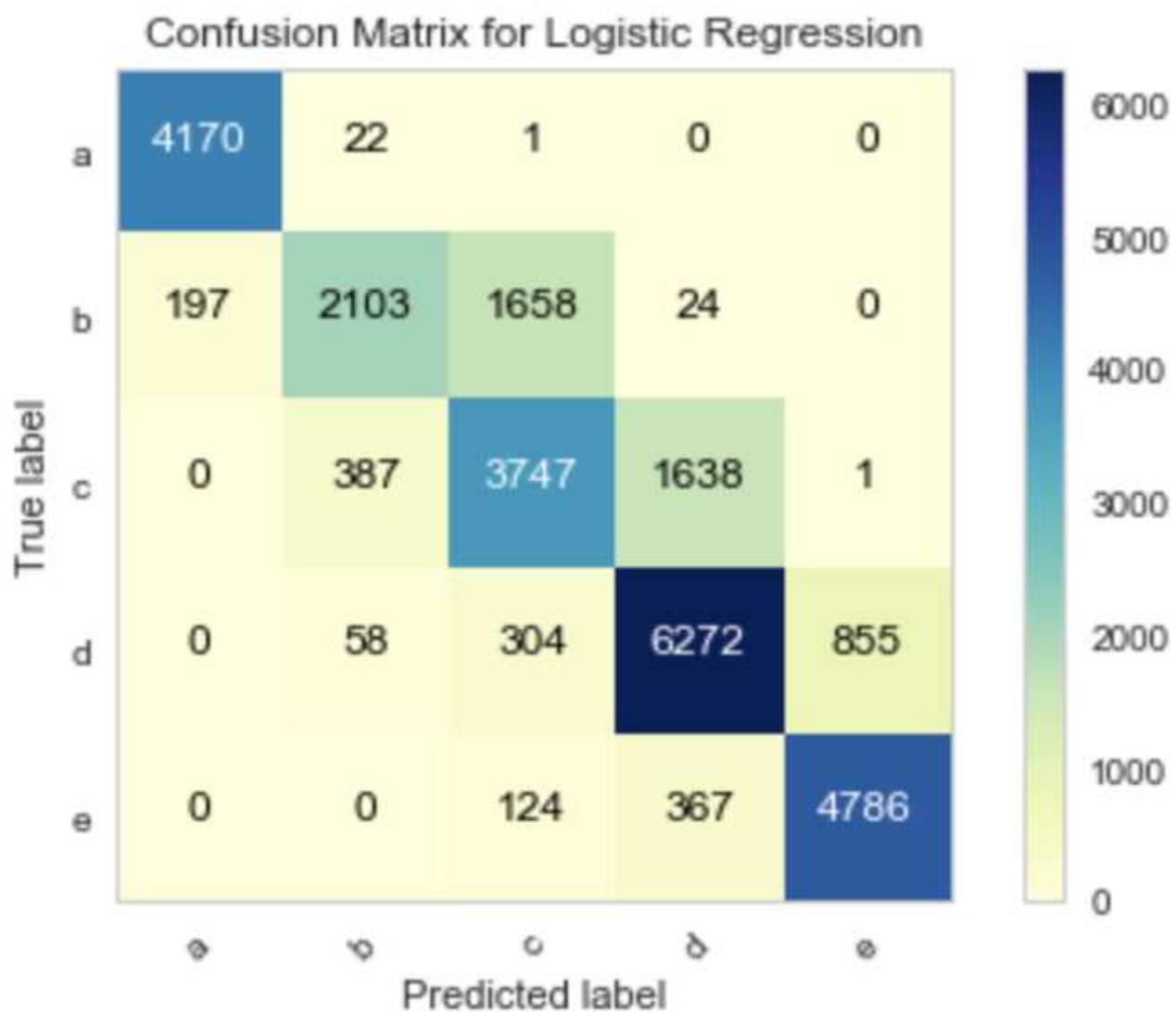


Figure 11: Confusion matrix for Logistic Regression [9]

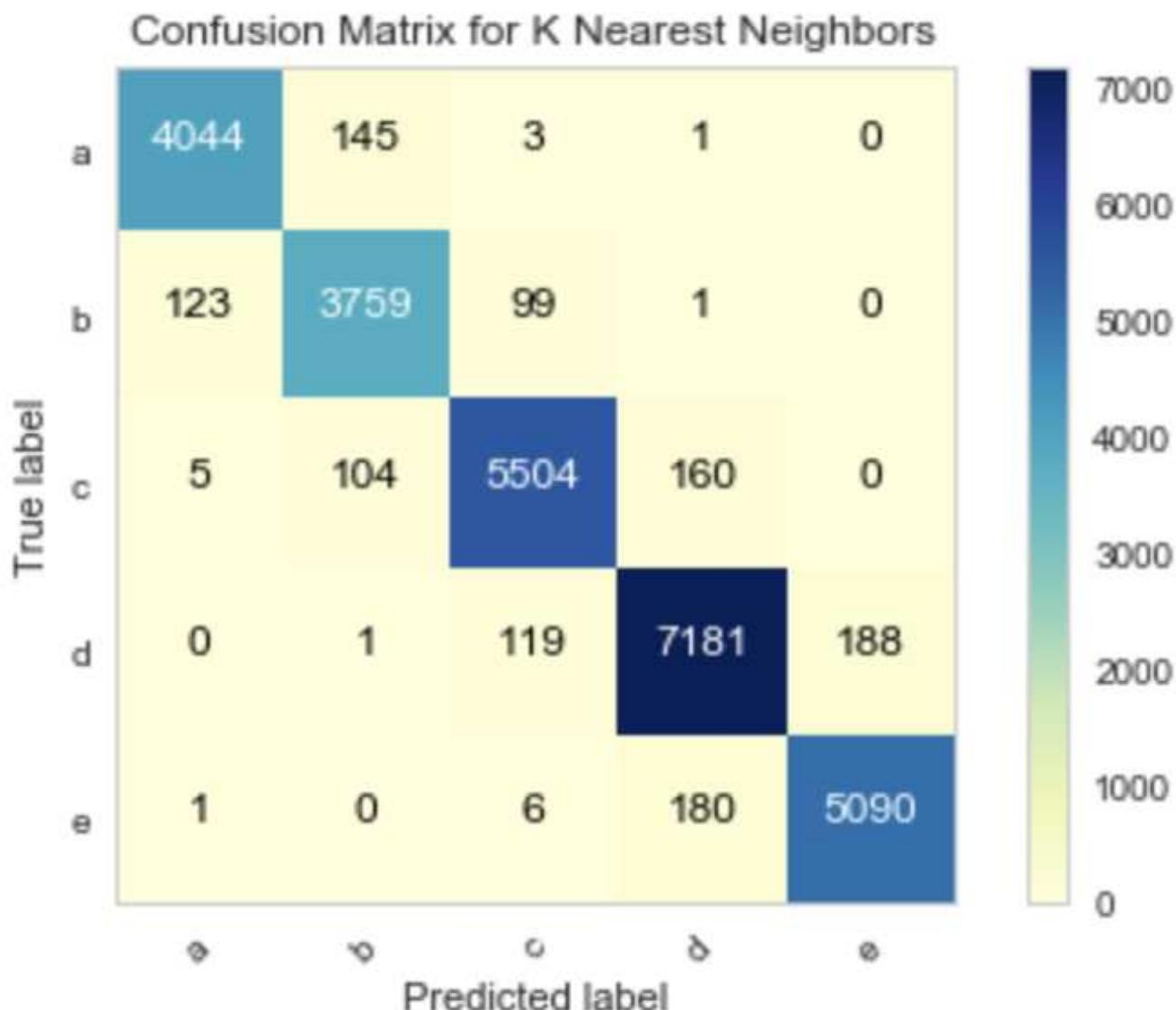


Figure 12: Confusion matrix for K Nearest Neighbors [9]

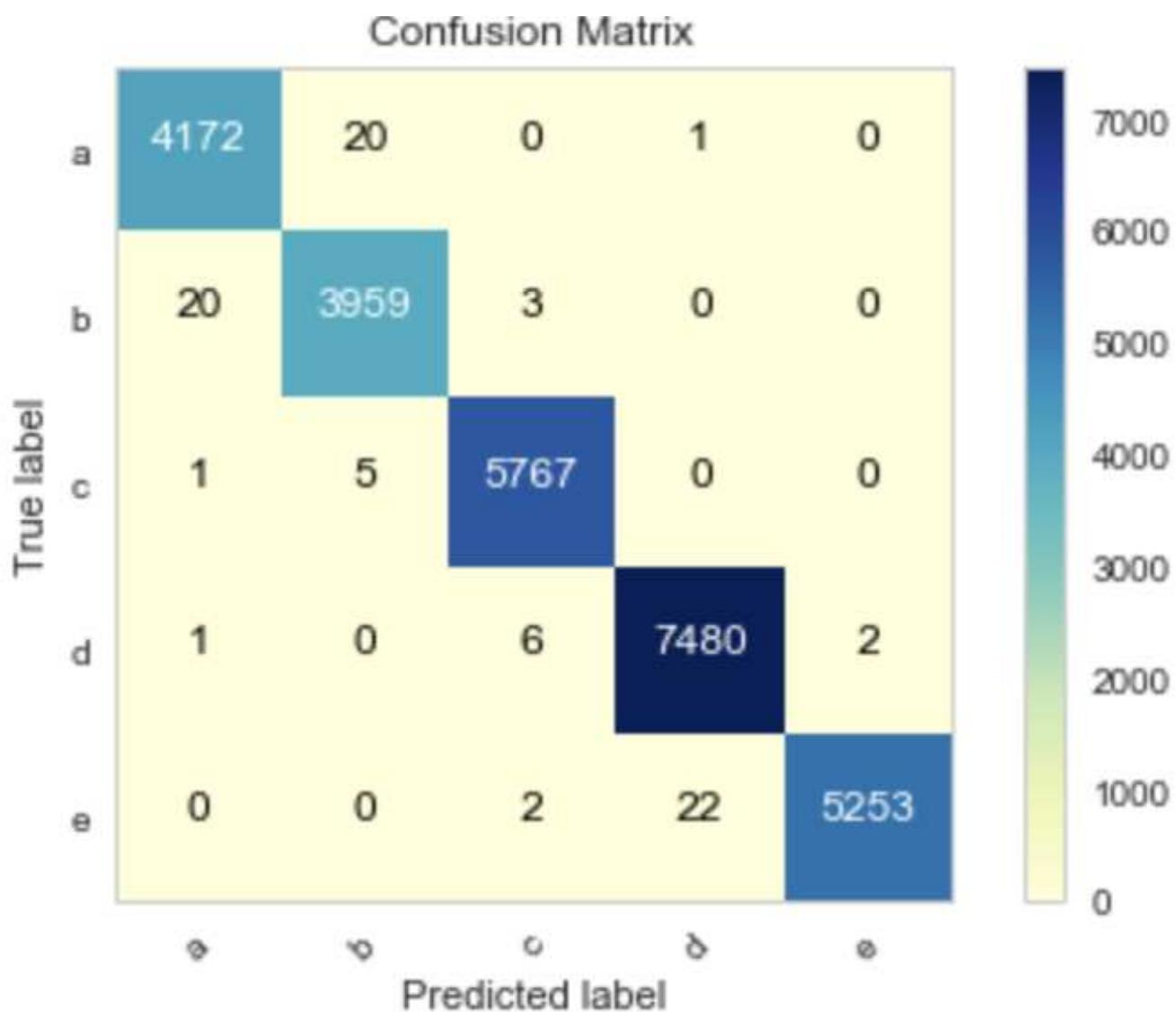


Figure 13: Confusion matrix for Random Forest [9]

```
bibtext report
```

```
=====
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
[2017-12-16 09.34.26] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.3s.
./README.yml
36:70     error    trailing spaces (trailing-spaces)
37:81     error    line too long (85 > 80 characters) (line-length)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Vegi, Karthik
hid: 231
paper1: Oct 29 17 100%
paper2: 100%
project: 100% Dec 4, 2017
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 231 project (null) 6226 report.tex
wc 231 project (null) 6232 report.pdf
wc 231 project (null) 250 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
71: We then display the top 5 countries as a pie-chart and the 5
countries are namely United States, France, Switzerland, Germany,
and Spain as shown in Figure \ref{fig:Fig7}.
```

```
73: \begin{figure}
```

```
74: \includegraphics[width=1.0\columnwidth]{images/fig7.png}
```

```
76: \label{fig:Fig7}
```

```
82: The top 5 countries with most fat content in the food items are
Serbia, United States, Switzerland, Germany, and Sweden as shown
in Figure \ref{fig:Fig8}.
```

```

84: \begin{figure}
85: \includegraphics[width=1.0\columnwidth]{images/fig8.png}
87: \label{fig:Fig8}
90: The top 5 countries with most saturated fat content in the food
    items are Serbia, United States, Germany, France and Switzerland
    as shown in Figure \ref{fig:Fig9}
92: \begin{figure}
93: \includegraphics[width=1.0\columnwidth]{images/fig9.png}
95: \label{fig:Fig9}
99: The top 5 countries with most trans-fat content in the food items
    are United States, Brazil, Canada, Australia, Russia, and Serbia
    as shown in Figure \ref{fig:Fig10}. \\
101: \begin{figure}
102: \includegraphics[width=1.0\columnwidth]{images/fig10.png}
104: \label{fig:Fig10}
107: The top 5 countries with most cholesterol content in the food
    items are United States, Canada, Portugal, Brazil, France, and
    Italy as shown in Figure \ref{fig:Fig11}. \\
109: \begin{figure}
110: \includegraphics[width=1.0\columnwidth]{images/fig11.png}
112: \label{fig:Fig11}
118: The top 5 countries with most sugar content in the food items are
    United States, Serbia, Switzerland, France, and Sweden as shown
    in Figure \ref{fig:Fig12}. \\
120: \begin{figure}
121: \includegraphics[width=1.0\columnwidth]{images/fig12.png}
123: \label{fig:Fig12}
126: The top 5 countries with most sodium content in the food items
    are United States, Hungary, Serbia, Sweden, and France as shown
    in Figure \ref{fig:Fig13}. \\
128: \begin{figure}
129: \includegraphics[width=1.0\columnwidth]{images/fig13.png}
131: \label{fig:Fig13}
165: The nearest neighbor puts each attribute list as a data point in
    the n-dimensional space, given n the number of attributes
    \cite{book-tan}. Once we have the training examples, we take each
    test example and compute its distance to the training example
    classes and assign a class label \cite{book-tan}. Any of the
    popular distance measures among Euclidean distance, Manhattan
    distance, Minkowski distance and Mahalanobis distance can be used
    \cite{book-tan}. The k denotes the k closest points to the test
    example \cite{book-tan}. Figure \ref{fig:Fig1} shows the
    algorithm \cite{book-tan}.
167: \begin{figure}
168: \includegraphics[width=1.0\columnwidth]{images/fig1.png}
170: \label{fig:Fig1}

```

```

245: \subsubsection{Bi-variate box-plots} Bi-variate box-plots go
beyond uni-variate box plots by showing the relationship between
the predictor variable and the response variable \cite{book-tan}.
We look at the bi-variate box-plots for each of the important
predictor variables namely, saturated fat, polyunsaturated fat,
sugars and salt and the response variable, nutrition grade.
Figure \ref{fig:Fig2} shows the bi-variate box plots. \\
247: \begin{figure}
248: \includegraphics[width=1.0\columnwidth]{images/fig2.png}
250: \label{fig:Fig2}
256: Correlation between data objects is the measure of the linear
relationship between the attributes of the object that are
continuous variables \cite{book-tan}. Correlation analysis is the
process of finding of the correlations between the different
predictor variables and identify high collinearity problem
\cite{book-shai}. The relationship could be either linear or non-
linear based on the given data \cite{book-tan}. The correlation
coefficient can range anywhere between -1 and 1, where 1
indicates a very high positive correlation and -1 indicates a
very high negative correlation \cite{book-shai}. Correlation plot
visually shows the correlation coefficient between the variables
in a nicely laid out plot. Figure \ref{fig:Fig3} shows the
correlation plot. \\
258: \begin{figure}
259: \includegraphics[width=1.0\columnwidth]{images/fig3.png}
261: \label{fig:Fig3}
306: \item Logistic Regression: With $l_2$ penalty, the accuracy of
logistic regression was 78.9\%. Figure \ref{fig:Fig4} shows the
confusion matrix. \\
308: \begin{figure}
309: \includegraphics[width=1.0\columnwidth]{images/fig4.png}
311: \label{fig:Fig4}
314: \item K Nearest Neighbors: With k as 3, the accuracy of kNN was
95.74\%. Figure \ref{fig:Fig5} shows the confusion matrix. \\
316: \begin{figure}
317: \includegraphics[width=1.0\columnwidth]{images/fig5.png}
319: \label{fig:Fig5}
322: \item Random Forest: With a number of trees as 100, the accuracy
of random forest classifier was 99.68\%. Figure \ref{fig:Fig6}
shows the confusion matrix. \\
324: \begin{figure}
325: \includegraphics[width=1.0\columnwidth]{images/fig6.png}
327: \label{fig:Fig6}

```

figures 13  
tables 0

```
includegraphics 13
labels 13
refs 13
floats 13
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8211
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# Recipe Ingredient Analysis

Sushant Athaley  
Indiana University  
sathaley@iu.edu

## ABSTRACT

Food is the unavoidable part of day to day of human life. Ingredients play a major role or are the basic requirement in preparation of any kind of food. We can find the humongous list of ingredients getting used across globally along with other details which constitute to big data. We explore ingredients getting used in various recipes across the globe to understand most used ingredient, key ingredients of various cuisine and the relationship between the ingredients to find out closely related ingredients which can always provide great dish if used together.

## KEYWORDS

i523, hid302, big data, ingredient, recipe, analysis, python, gephi

## 1 INTRODUCTION

Ingredients are vital for human existence as well as for food or restaurant industry. We use it every day for cooking and food industry uses it to produce consumable for their customers. Ingredient inspires chefs to come up with new culinary artistry. So what do we know about this essential element of the life and what data tell us? Ingredients come in different size, color, shape, flavor, nutrition, taste, texture, grows in specific weather conditions and this provides a great opportunity for various analysis which can be useful for the human being as well as business industries. There can be multiple analysis carried out on the ingredients but main focus of this study is on the ingredients used in various recipes across the cuisines understand most used ingredients, key cuisine ingredients and ingredient relationship.

This study is organized as follows, section *Big Data and Food* touch open big data and its application in food industry, section *Ingredient* defines ingredient and it's various characteristics. section *Ingredient Analytics and Related Studies* describes various analytics which can be performed on the ingredient with some examples and studies. Section *Project* describes the aim of this study. Section *technologies* provides information on the tools and technologies used for this project. Section *Methodology* covers overall process carried out in this project. Section *Dataset* describes data structure used along with loading process and data findings. Section *Analysis and Findings* describes various analysis carried out on the data and the visual representation of the analysis. Section *Shortcomings* captures shortcomings of the project. Section *Limitations* talks about limitations and what else can be done with this dataset which is not covered in the current scope of the project. Section *Conclusion* concludes the study.

## 2 BIG DATA AND FOOD

Big Data is defined in lot many different ways but one of the interesting ways it has been defined is in terms of three V's which are Volume, Velocity, and Variety. Big data is generated in great volume

typically in the gigabyte or more which makes data processing difficult. Data *velocity* has been increased due to the real-time data streaming from various applications like social media or different type of sensors recording data continuously. Big data comes in *variety* of format like structured or unstructured data. Data varies in various format like text, pictures, audio, videos, 3D, social media and so on. These big data characteristics pose challenges in terms of overall data lifecycle management. Some of the examples of big data usage are the recommendation service, predictive analytics, data analytics, pattern identification, and machine learning.

Over the period of time, food has grown from basic necessity to big food industry. Food industry covers lot many businesses under its umbrella like agriculture which is growing/raising/catching food, food production or manufacturing, food processing, food safety and compliance, distribution, marketing, food retailing and food service [9]. Ultimately this wide array provides us with a huge opportunity for big data application in food industry.

Agriculture is moved on to precision agriculture with the rise of new technologies. Precision agriculture is a practice of farming more accurate and controlled when it comes to the growing of crops and raising livestock. A key component of this farm management approach is the use of information technology and a wide array of items such as GPS guidance, control systems, sensors, robotics, drones, autonomous vehicles, variable rate technology, GPS-based soil sampling, automated hardware, telematics, and software. Big data gathered by these technologies are used to guide both immediate and future decisions on when it is best to apply chemical, fertilizer or seed [19].

The distribution includes all activities of moving food from food producer to the consumer. Big data can provide valuable analytics in determining the best transportation methods and routes. By analyzing transport methods and making them more efficient, spoilage and damage can be reduced, allowing a greater percentage of products to make it from the farm to the customer. This is important as lot many time food contains perishable items which can result in bio-waste if not handled properly during the transportation. Reducing this waste will increase profits, as well as the amount of food produced, and will have a positive impact on the environment [17].

Food safety is another growing concern in the food industry as it has direct implication to the human health. Analyzing data about food quality helps to detect spoiled food, preventing it from reaching the customer. This analysis can also help producers and distributors in the food industry identify contaminated food, and isolate its source and current location. Not only does this allow for faster recalls, minimizing the number of people exposed to the food, it also allows for targeted recalls rather than blanket ones. This saves the company significant amounts of money, as fewer items need to be removed from shelves and replaced [17].

Big data is allowing restaurant chains to closely monitor every aspect of their business. By collecting information from every

individual restaurant, it is possible for food analytics to detect patterns such as what menu items perform best in which regions, how much food needs to be stocked and prepared for a given week or even a particular time of day, and what building layout provides the best and most efficient experience [17]. Sentiment analysis can help in understanding the customer emotions. Preventive action can be taken to address the customer dissatisfaction.

### 3 INGREDIENT

Food is defined as “Edible or potable substance (usually of animal or plant origin), consisting of nourishing and nutritive components such as carbohydrates, fats, proteins, essential mineral and vitamins, which (when ingested and assimilated through digestion) sustains life, generates energy, and provides growth, maintenance, and health of the body” [4]. Thus food is the basic necessity for human for the sustainability. Food can be eaten raw, cooked or processed. As human race evolved over the period of time, the way we eat food is also evolved. Food cooking is just not the basic necessity but its an art and science in today’s era. Food preparation consists of various cooking techniques, tools, and ingredients to make it palatable or edible by humans. The ingredient is by far the most important part of any food or recipe preparation. The recipe consists of the list of ingredients and the set of instruction to cook particular food dish [8]. An ingredient is defined as “Any of the foods or substances that are combined to make a particular dish” [16]. Ingredients impart various flavors, aroma, texture, and color to the cooking dish. Ingredients are mostly derived from vegetables, fruits, nuts, grains, living organisms, herbs, flowers, and spices. It comes in both solid and liquid forms. Another characteristic of ingredients is the nutritional value they provide which is essential for the human body.

### 4 INGREDIENT ANALYTICS AND RELATED STUDIES

Ingredients characteristics and the combination of other related data provides various opportunities to analyze ingredient in different ways. Flavor network and the principle of food pairing by Yong-Yeol Ahn et al. [1] is the most referenced study in terms of ingredient analysis. They built a bipartite network consisting of ingredients and flavor compounds imparted by those ingredients. This flavor network connects two ingredients if there is at least one flavor compound is shared by those ingredients. More the flavor compound ingredient they share more strongly they are related. This network revealed that fruits and dairy products are close to alcoholic drinks, and mushrooms appear isolated, as they share a statistically significant number of flavor compounds only with other mushrooms. They further studied food pairing hypothesis and found out that in North American recipes, the more compounds are shared by two ingredients, the more likely they appear in recipes. By contrast, in East Asian cuisine the more flavor compounds two ingredients share, the less likely they are used together. Analysis of the flavors present in ingredient can provide us with the categorization of the different ingredient by the flavor profile which can be helpful in deciding substitute ingredient if a certain ingredient is not present or pairing ingredient from different flavor categories

to construct the dish as per the taste required. This analysis also helps to understand which ingredients cannot be used together.

Another analysis is carried out to correlate ingredient across recipes to come up with top 50 combinations of ingredients which can be used together [11]. Some of the combinations finding from this study are interesting and fun to experiment

- tomato, garlic, oregano, onion, basil
- vanilla, cream, almond, coconut, oat
- onion, black pepper, vegetable oil, bell pepper, garlic
- cumin, coriander, turmeric, fenugreek, lemongrass

Flavourspace application provides functionality to search recipe based on the ingredients, suggests alternate ingredient if not present, adjust the recipe as per the taste which is a good example of big data analytics in food industry [18].

Foodpairing application takes another approach to form the connection between unfamiliar ingredients and provides information on how to use such ingredient to make a dish, this is very helpful in terms of sustainability as we can use ingredient which is ample available but not in use due to the absence of information on using such ingredients [14].

Recipe recommendation system uses users recipe browsing history or rating history to suggest the recipe. It also relay on the ingredient present in the recipe and look for the overlap or key ingredients while matching other recipes. Another approach is to recommend recipe based on the nutritional values or healthy food choice which is dependent on the ingredient used in the recipe. Models are made to recommend recipe based on the available ingredients and personal nutrition needs. Chen-Yuen et al. [6] derived network of complimenting and substituting ingredients. They also demonstrated that network can be used to predict which recipe would be successful. To understand the complimenting ingredient they constructed network based on pointwise mutual information (PMI) defined on pairs of ingredients. This PMI provides the probability of those two ingredients occurring together. Their study found out 2 main cluster as savory and sweet dishes along with the a satellite cluster of mixed drink ingredients. This study also finds out ingredient adjustment and substitution based on the comments on the recipe. Recipe comments provides insight into which ingredients quantity is increased or decreased to get more flavors or which ingredient is used instead of some ingredient in the recipe since ingredient mentioned in recipe is not present or to get different taste. The words like add, omit, instead, adding, using, more etc in comments provides this insight. Ingredient which are considered as unhealthy like sugar, fats are often reduced and ingredient which adds flavors like soy sauce, lemon juice, cinnamon are added more in quantity. Chicken can be substituted by turkey, beef, sausage, chicken breast, bacon and olive oil by butter, apple sauce, oil, banana, margarine, and Tilapia by cod, catfish, flounder, halibut, orange roughy to name few.

The researcher at IBM have built a program that uses math, chemistry, and vast quantities of data to churn out new and unusual recipes. The new recipes are generated by ‘mutating’ the ingredients of existing recipes, and then fusing these with other recipes, resulting in all sorts of new hybrid concoctions. This idea, known as a genetic algorithm, is modeled after the process of genetic change [3].

Another study conducted on most used ingredient provides insight that sugar, oil, pepper, and salt are most commonly occurring ingredient, among spices clove, in vegetable onion, garlic , and tomatoes, butter in milk product, eggs followed by chicken in the animal product are the most used ingredient in the categories [5]. This information can help in better planning and sourcing of such ingredients which are in high demand.

Ingredient nutrition analysis can help find out nutrition of the food prepared by those ingredients. This would be helpful in menu planning where nutrition information is the key factor such as school, hospitals or any other dietary program [7].

Yannick Kimmel [13] analyzed top 20 recipes on allrecipes.com website for last 20 year to understand the food trends in the USA. Word cloud analysis on recipe title revealed that cookie, chicken, chocolate, banana, salad, bread, potato, pie, cake, and bake are most frequently used in the top recipe titles. The ingredient word cloud includes sugar, white, ground, butter, salt, bake, and chop which are fundamental words in cooking. Recipe calorie analysis shows that there is jump in calories in initial year but it drop slowly over period of time which can be reflection of focus on healthy food. Analysis also reveals that there is increase in usage of olive oil which might be the result of health benefits provided by olive oil.

Recipe cost is calculated by including the cost of the ingredient used in that recipe. Ingredient cost as per the quantity used in recipe provides base information to calculate the price of any recipe. This ingredient cost analysis provides an avenue to reduce the cost of the recipe by using substitute ingredient of lesser cost. This can also help in household budget to keep in check as well as make restaurant industry profitable.

Ingredient used in recipe can provide insight into type of weather received by that cuisine as ingredient can grow in certain weather condition. This can help chef locally source the ingredient and maintain local agriculture sustainability.

According to study food also contains medicinal properties and can be used for the healing. Traditional healing methods like *Ayurveda* in India and Chines medicine relay on food or various herbs medicinal properties for the healing. Ingredients has medicinal properties like Decreasing and Controlling Inflammation, Balancing Hormones, Alkalizing the Body, Balancing Blood Glucose, Detoxifying and Eliminating Toxins and Improving Absorption of Nutrients. Green vegetables like kale, wheat grass and spinach, sea vegetables are considered some of the healthiest foods and known to help slow aging [2]. Analysis of food ingredients for the medicinal properties to classify those food with various medicinal values and effect on human body can greatly help as this treatment can be low cost as compare to other medical treatments.

We also found ingredient relationship analysis and network graph generated in another study [20]. This analysis shows ingredient cluster of 5 cuisines.

## 5 PROJECT

This project study is conducted to analyze ingredients getting used in various recipes across the cuisines to find out

- Most used ingredients across cuisines or globally
- Key ingredients used by cuisines

- Ingredient relationship to understand the related ingredients and provide complimentary ingredient network

### 5.1 Technologies

Technologies and tools used in this projects are

- Python version 3.6 is used for data load and processing
- Gephi 0.9.2 for visualization
- Spyder 3.0 as a Python IDE

### 5.2 Code Organization

Code is checked-in in Github at location

<https://github.com/bigdata-i523/hid302/tree/master/project/code>

Code is organized as described in Figure 1

[Figure 1 about here.]

#### Python Scripts

- *ingredientAnalysis.py* - This python script loads dataset from datafile train.json present in *data* directory and process dataset to find out recipe distribution across cuisines, top 20 ingredient used across cuisines and top 10 key ingredients for every cuisine. The graph generated during analysis is stored in *images* folder. This codes inspiration can be found out at Kaggle's What Cooking competition which we modified as per our project need [15], [10].
- *ingCluster.py* - loads dataset from datafile train.json present in *data* directory and process dataset to create relationship file required by Gephi in excel format. It establishes ingredients relationship by relating ingredient in recipe with each other to generate *nodes.xlsx* and *edges.xlsx* files and stores in *data* directory. These generated files are then imported into the Gephi to create the visualization.
- *geph Ing\_big\_data.gephi*
  - This is project file from Gephi which can be re-opened in Gephi software to view or re-run the analysis.

### 5.3 Methodology

We followed methodology as described to complete our study

- *Identify Data Source* - We analyzed various sources of ingredients and finalized the data source
- *Collect/Extract Data* - We analyzed various ways of extracting data from the data source and finalized our approach on data extraction process
- *Load Data* - Load data using Python script for the analysis
- *Clean/Filter Data* - Process loaded data for the clean up to avoid unwanted data
- *Process Data* - Process cleaned up data through python scripts to analyze most used ingredient, ingredient distribution across cuisines and per cuisine
- *Generate Ingredient Relationship Network* - Gephi software is used to analyze the relationship and to find out the ingredient modularity. We investigated what kind of data is needed for Gephi for the analysis and we understood that Gephi needs node and edges which is nothing but the relationship between the nodes. Node contains node id and edges contains source node id and target node id which depicts source node is related to target node and this file can

be generated in excel file format. Python script is used to create the network files required by the Gephi tool. Python script generated Node and Edges file in excel format so that it can be imported into Gephi. Distinct ingredients used in recipes becomes the nodes. Edges or relationship between ingredients is derived by relating ingredients appearing in the same recipe. All ingredient in the same recipe is considered related to each other.

- *Import Data to Gephi* - Network files created by Python are imported in Gephi to produce the graph for the visualization.
- *Clean/Filter Data in Gephi* - Gephi tools data laboratory is used to clean up the data and filters are applied to provide usable network visualization.
- *Data Processing in Gephi* - Process data in Gephi by applying layouts and statistics to generate the graph
- *Visualize and Publish Results* - Gephi tools data laboratory is used to clean up the data and filters are applied to provide usable network visualization.

Figure 2 shows pictorial representation of the methodology used for this project to analyze ingredient data.

[Figure 2 about here.]

#### 5.4 Data Gathering

The first step was to source the data. We were interested in the dataset which provides recipe information along with the ingredient used in the recipe. Since we wanted to analyze distribution across cuisines, data should also contain cuisine tagging. We evaluated 2 ways of gathering the data, generate the data ourselves or use publicly available data.

The dataset can be generated by pulling recipe data from various online applications or pick from publicly available datasets. There are lot of applications online like allrecipes, Food, Yummly etc which hosts thousands of recipes and not to forget about recipe site available in every country, if we consider all these sources then it can easily contribute to huge dataset. Yannick Kimmel [13] demonstrated in his recipe analysis project how recipe data can be source directly from the application. He did analysis of top 20 recipes from allrecipes website which is the largest web application hosting the recipes. He used Selenium package in Python to scrap allrecipes which can handle AJAX used in the application. Each recipe in allrecipe can be identified using unique identifier and follows generic format as [allrecipes.com/recipe/\[Unique ID number\]](http://allrecipes.com/recipe/[Unique ID number]). This generic URL is used by passing different unique id number to retrieve the recipe page and then find-element method is used to read various attributes like title, rating, reviews, calories per serving, prep time, cook time, total time and ingredients. We can follow the same approach to generate the dataset from different recipe sites for our analysis but we finalized publicly available dataset at Kaggle application satisfying need for this project to save the time.

#### 5.5 Dataset

The dataset for this study is sourced from Kaggle application [12]. This dataset is publicly available and featured in *What's Cooking?* competition. This dataset is provided to Kaggle by Yummly which is the application which hosts recipes online. This dataset is in JSON

format and of 12MB size. This dataset contains recipe id, cuisine and list of ingredients as described in Figure 3.

[Figure 3 about here.]

This dataset contains total 39774 recipes across various cuisines. We used two different methods to load this data. Cuisine and ingredient analysis is done by loading data into *pandas dataframe* and to analyze ingredient relationship data has been loaded into *json* object. Figure 4 shows the code for data loading used in this project.

[Figure 4 about here.]

Ingredient extraction from the data structure and processing was challenging as ingredients are listed comma separated for each recipe. Also, ingredient list can vary by recipe and there is no proper structure. We observed shortcoming of dataset as

- *Ingredient Duplication* - ingredient appears in the ingredient list in various forms but it's the same ingredient which gives duplicate data. For example, salt appears as salt, kosher salt, Morton Salt, sea salt, table salt, Himalayan salt, fine sea salt, low sodium salt, fine salt. This is the same ingredient but come across in recipe as a different ingredient and getting counted as a separate ingredient in the analysis
- *Ingredient Name along with measure* - ingredients are listed along with measures like (10 oz.) frozen chopped spinach, (10 oz.) frozen chopped spinach, thawed and squeezed dry, (14.5 oz.) diced tomatoes and getting counted as a separate ingredient
- *Branded Ingredients* - ingredients are listed along with the brand name like KRAFT Reduced Fat Shredded Mozzarella Cheese, Johnsonville Smoked Sausage, Johnsonville Mild Italian Sausage Links etc and also constitutes to the ingredient list
- *Country Name with Ingredients* - ingredients are listed along with the country name like japanese cucumber, korean chile paste, Japanese soy sauce etc and also constitutes to the ingredient list
- *Ingredient name too long* - some ingredient name are too long to be an ingredient like *wish-bone light asian sesame ginger vinaigrette dressing*
- *Ingredient with application* - ingredient names like chopped onion, diced tomatoes, chopped cilantro, grass-fed beef, grass-fed butter, cut up chicken

This variation makes difficult to get the proper ingredient list for the analysis. Extensive work is needed to clean and correct the noisy data so that proper analysis can be carried out. This correction process is not carried out as part of this project.

Certain ingredients like salt or water etc should be avoided from the analysis as those are not the ingredient we are looking for the analysis. We tried to clean such elements during ingredient relationship analysis but we had little success as those ingredients are present in the dataset in various forms.

## 5.6 Analysis and Findings

**5.6.1 Recipe Distribution By Cuisine.** We first analyze entire dataset to understand the total number of recipes and their distribution across various cuisines. We use Pythons Panda library to get the total recipe count as 39774 and plot the distribution. Figure 5 shows number of recipes per cuisine. Our observations from this analysis are

- Dataset is heavily dominated by Italian cuisine followed by Mexican cuisine which shows popularity of those cuisines
- Very fewer recipes from Russian and Brazilian cuisines which shows very less contribution from those areas
- No recipes from some regions like Germany, Canada which might be due to the recipes are not uploaded by users from that regions
- This also highlights another shortcoming of the dataset that it doesn't have equal representation of all cuisines which might give us biased analysis

[Figure 5 about here.]

Table1 describes recipe count for every cuisine.

[Table 1 about here.]

**5.6.2 Most Used Ingredients All Cuisines.** The second analysis is carried out to understand top 20 ingredients getting used across cuisine or globally. As per our study most used distinct ingredients across cuisines in order are

- Salt
- Olive Oil
- Onions
- Water
- Garlic
- Sugar
- Butter
- Black Paper
- All-purpose flour
- Vegetable Oil
- Eggs
- Soy Sauce
- Green Onions
- Tomatoes
- Carrots

Ingredient *Salt* is obvious topper followed by *Oil* and *Onions*. This also proves our craving for salty and fatty food. Top 20 ingredient also contain duplicate ingredient like garlic and garlic clove, salt and kosher salt, eggs and large eggs which shows shortcoming of the dataset. Also ingredient like salt, oil and water could be avoided to get analysis of real ingredients as these are commonly use ingredient and doesn't contribute much to the study. Figure 6 shows top 20 ingredient across cuisines.

[Figure 6 about here.]

**5.6.3 Ingredients Distribution By Cuisines.** The third analysis is carried out to understand key ingredient for each cuisine. These key ingredients define those cuisines and provide unique test characterized by that cuisine. We limited ingredient list to top 10 to get the key ingredients for each cuisine. Study shows key ingredient for our top 5 cuisines as follows

- *Italian* - Olive oil, garlic, cheese, black pepper, onion and butter
- *Mexican* - onion, cumin, garlic, chili powder, jalapeno chilies, sour cream, tortillas and avocado
- *Southern US* - butter, all-purpose flour, sugar, eggs, baking powder, milk and butter milk
- *Indian* - onion, garam masala, turmeric, garlic, cumin and oil
- *Chinese* - soy sauce, sesame oil, corn starch, sugar, garlic, green onions and scallions

Similarly we show key ingredient of all other cuisines present in the dataset and we observe that it is very close representation of all cuisines. Figure 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 shows top 10 key ingredient used in the corresponding cuisines.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

[Figure 18 about here.]

[Figure 19 about here.]

[Figure 20 about here.]

[Figure 21 about here.]

[Figure 22 about here.]

[Figure 23 about here.]

[Figure 24 about here.]

[Figure 25 about here.]

[Figure 26 about here.]

**5.6.4 Ingredients Relationship.** Forth analysis is carried out to understand the relationship between the ingredient to find out ingredient clusters. This analysis helps us understand the ingredient combinations which can be used together to provide great dish every time. This model can be used to predict ingredients for certain recipe based on the cluster. We used Gephi tool to analyze and produce the graph for this analysis. Gephi accepts network structure in terms of Node and Edge relationship. We created this network using python by relating all ingredients present in the recipe with each other. Ingredients become the node and source and target nodes become the edges. These network files generated in excel spreadsheet and converted to CSV format and imported into the Gephi tool. Import created 5405 Nodes and 290828 edges for processing and analysis. Force Atlas 2 layout present in Gephi has been applied to the network which brings nodes with higher weights and shared connections closer to each other. We also used Gephi Data Laboratory to clean up duplicate or unwanted nodes.

Filtering based on Degree Range and Edge Weight has been applied to data to reduce node and edges to get the graph which can be used for analysis and avoid crashing Gephi due to large data. Modularity statistic uncovered 5 ingredient clusters which can be identified by different colors in the graph. This cluster can approximately relate to the cuisines present in our dataset and confirms our earlier analysis of ingredient by cuisine.

- Orange - Mexican
- Brown - Indian
- Blue - Chinese
- Green - Italian
- Gray - Southern US

This analysis also provides us with the complimentary ingredient network which can be used together to construct tasty dish. We can see that there are two type of combination one is savoury and another is sweet. The complimentary ingredients as per our study are

- Sugar, butter, all-purpose flour, large eggs, heavy cream, baking powder, cinnamon, flour, lemon, vanilla
- Olive oil, garlic, black paper, onion, cheese, basil, parsley, oregano, white wine, shallots, lemon juice, bell paper
- Onion, garlic, pepper, tomato, bay leaves, paprika, potatoes, chicken, shrimp, celery, green paper, garlic powder, dried thyme
- Tomato, ground cumin, chicken, cilantro, jalapeno chilies, ground beef, sour cream, chili powder, avocado, corn tortillas, black beans, salsa, lime, green chilies, oil, turmeric, garam masala, coconut milk
- Oil, green onions, scallions, carrots, garlic, ginger, fish sauce, soy sauce, rice vinegar, sesame oil, corn starch, brown sugar, honey

Graph also shows overlap between following ingredients which confirms that those are the commonly together used ingredients in the recipes. We observe those combination in Indian and Italian cuisine.

- Onion, garlic
- Olive oil, black paper

Figure 27 shows ingredient cluster of more than 1000 nodes. This graph is nice to look at but difficult to read due to lot many nodes and edges in the graph.

[Figure 27 about here.]

Figure 28 shows ingredient cluster of around 100 nodes. We generated this graph by reducing nodes and edges to make it more readable. This graph provides us with our top 5 cuisine clusters.

[Figure 28 about here.]

## 5.7 Shortcomings

Improper documentation of ingredient names in the dataset reduces the correctness of this analysis. In absence of proper ingredient name and duplication of ingredient name prevents getting exact ingredient weight into the analysis. A dataset with uniform ingredient name can help this analysis to achieve its best. If we don't find proper ingredient name then this analysis needs to include extensive data cleaning process which can be considered an improvement to this project.

Network file creation algorithm can be enhanced further by considering the number of recipes for the ingredient to provide additional weight to the relationship which can provide the stronger bond between the ingredients.

## 5.8 Limitations

This dataset can be analyzed to find out ingredient overlap between various cuisine and can provide insight into the influence of one cuisine on another which is not covered as part of this study. Usually, geographically neighboring cuisines are influenced by each other as they share common ingredients.

## 6 CONCLUSION

This project shows most used ingredient, ingredient distribution by cuisine and predictive ingredient relationship model as per the goal of the project. We also show various opportunities present with ingredient data analysis and role of big data analytics. We prove human craving for salty and fatty food as salt and oil are most used ingredient across cuisines as per the analysis. We understand now based on our analysis key ingredient of any cuisine. Ingredient cluster shows why those ingredients are the base of certain cuisine and recipe of those ingredients always turn out delicious. We also crave for the good data so that we can provide more accurate analysis of the ingredients. Ingredient analysis has potential not only to help restaurant and food industry but it can help with our social responsibility of sustainability and understanding different cuisines and culture. As food industries interest grows in big data analytics, we will continue to see more evaluations of the ingredients.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions in this project. The author would also like to acknowledge Kaggle application for hosting ingredient dataset which is used in this project and various application users contributing in data analysis. We also acknowledge various online resources which helped understand Python and Gephi.

## REFERENCES

- [1] Bagrow James P Ahn Yong-Yeol, Ahnert Sebastian E. 2011. Flavor network and the principles of food pairing. (2011). <https://www.nature.com/articles/srep00196#supplementary-information>
- [2] Dr. Axe. 2017. Food Is Medicine. web. (2017). <https://draxe.com/food-is-medicine/>
- [3] Aatish Bhatia. 2013. A New Kind of Food Science: How IBM Is Using Big Data to Invent Creative Recipes. web. (2013). <https://www.wired.com/2013/11/a-new-kind-of-food-science/>
- [4] businessdictionary. 2017. Food. web. (2017). <http://www.businessdictionary.com/definition/food.html>
- [5] Usashi Chatterjee, Vinit Kumar, and Devika P. Madalli. 2016. Formalizing Food Ingredients for Data Analysis and Knowledge Organization. *COLLNET Journal of Scientometrics and Information Management* 10 (07 2016), 289–309. [https://www.researchgate.net/publication/311337510/Formalizing\\_Food\\_Ingredients\\_for\\_Data\\_Analysis\\_and\\_Knowledge\\_Organization](https://www.researchgate.net/publication/311337510/Formalizing_Food_Ingredients_for_Data_Analysis_and_Knowledge_Organization)
- [6] Lada A Adamic Chun-Yuen Teng, Yu-Ru Lin. 2011. Recipe recommendation using ingredient networks. web. (2011). <https://arxiv.org/pdf/1111.3919.pdf>
- [7] S. M. Church. 2015. The importance of food composition data in recipe analysis. web. (2015). <http://onlinelibrary.wiley.com/doi/10.1111/nbu.12125/abstract>
- [8] collinsdictionary. 2017. Recipe. web. (2017). <https://www.collinsdictionary.com/us/dictionary/english/recipe>
- [9] FOODINDUSTRY. 2017. FoodIndustry.Com Business Categories. web. (2017). <https://www.foodindustry.com/features/food-industry-businesses/>
- [10] Froll. 2015. 10 Most Used Ingredients by cuisines. web. (2015). <https://www.kaggle.com/mrfroll/10-most-used-ingredients-by-cuisines>

- [11] inkhorn82. 2014. A Delicious Analysis. web. (2014). <https://www.r-bloggers.com/a-delicious-analysis-aka-topic-modelling-using-recipes/>
- [12] kaggle. 2015. What's Cooking? web. (2015). <https://www.kaggle.com/c/whats-cooking/data>
- [13] Yannick Kimmel. 2016. All the recipes: Scraping the top 20 recipes of all-recipes. web. (May 2016). <https://nycdatascience.com/blog/student-works/recipes-scraping-top-20-recipes-allrecipes/>
- [14] Bernard Lahousse. 2016. Using Big Data to Transform Unfamiliar Ingredients Into Tasty Recipes. web. (2016). <https://foodtechconnect.com/2016/04/20/big-food-data-recipes-from-unfamiliar-ingredients/>
- [15] Manuel. 2015. 10 Most Used Ingredients. web. (2015). <https://www.kaggle.com/manuelatadvice/nomame>
- [16] oxforddictionaries. 2017. Ingredient. web. (2017). <https://en.oxforddictionaries.com/definition/ingredient>
- [17] QUANTZIG. 2017. HOW BIG DATA IS REVOLUTIONIZING FOOD INDUSTRY PRACTICES. web. (2017). <https://www.quantzig.com/blog/big-data-revolutionizing-food-industry-practices>
- [18] Matthew Robinson. 2015. Big Data Analytics and Food Come Together At Flavourspace. web. (2015). <http://www.theculinaryexchange.com/food-innovation/big-data-analytics-and-food-come-together-at-flavourspace/>
- [19] REMI SCHMALTZ. 2017. What is Precision Agriculture. web. (2017). <https://agfundernews.com/what-is-precision-agriculture.html>
- [20] Davide Totaro. 2017. foodgraph. web. (2017). <https://github.com/d-t/foodgraph>

## LIST OF FIGURES

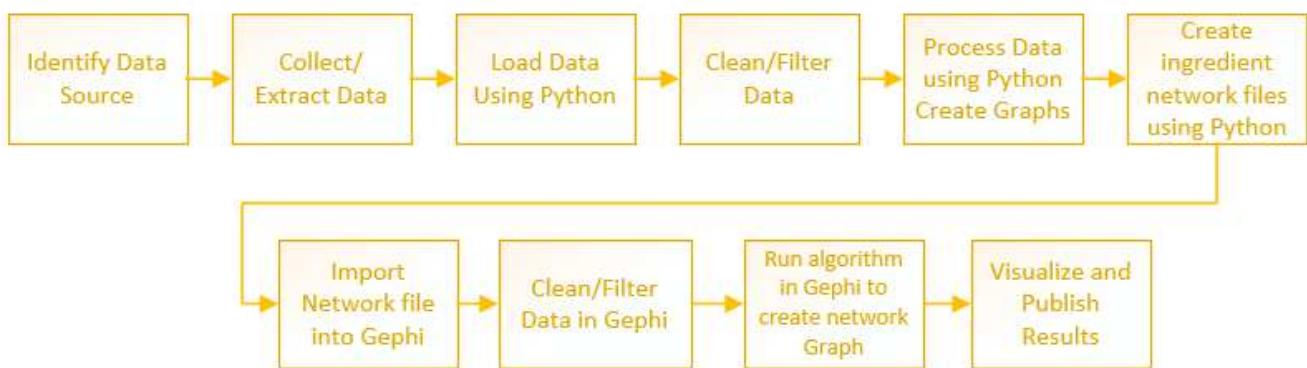
1	Code Structure	9
2	Flowchart of the Methodology to Analyze Ingredients	9
3	Ingredient Data Structure	9
4	Data Loading	10
5	Recipe Distribution By Cuisine	10
6	Top 20 Ingredients	11
7	Top 10 Ingredients	12
8	Top 10 Ingredients	13
9	Top 10 Ingredients	14
10	Top 10 Ingredients	15
11	Top 10 Ingredients	16
12	Top 10 Ingredients	17
13	Top 10 Ingredients	18
14	Top 10 Ingredients	19
15	Top 10 Ingredients	20
16	Top 10 Ingredients	21
17	Top 10 Ingredients	22
18	Top 10 Ingredients	23
19	Top 10 Ingredients	24
20	Top 10 Ingredients	25
21	Top 10 Ingredients	26
22	Top 10 Ingredients	27
23	Top 10 Ingredients	28
24	Top 10 Ingredients	29
25	Top 10 Ingredients	30
26	Top 10 Ingredients	31
27	Ingredient Cluster	32
28	ingredient Cluster 100 Nodes	33

```

code
- ingredientAnalysis.py
- ingredientAnalysis.py
- data
  - train.json
  - nodes.xlsx
  - edges.xlsx
- images
- gephi
  - geph Ing big data.gephi

```

**Figure 1: Code Structure**



**Figure 2: Flowchart of the Methodology to Analyze Ingredients**

```

{
  "id": 24717,
  "cuisine": "indian",
  "ingredients": [
    "tumeric",
    "vegetable stock",
    "tomatoes",
    "garam masala",
    "naan",
    "red lentils",
    "red chili peppers",
    "onions",
    "spinach",
    "sweet potatoes"
  ]
},

```

**Figure 3: Ingredient Data Structure**

```

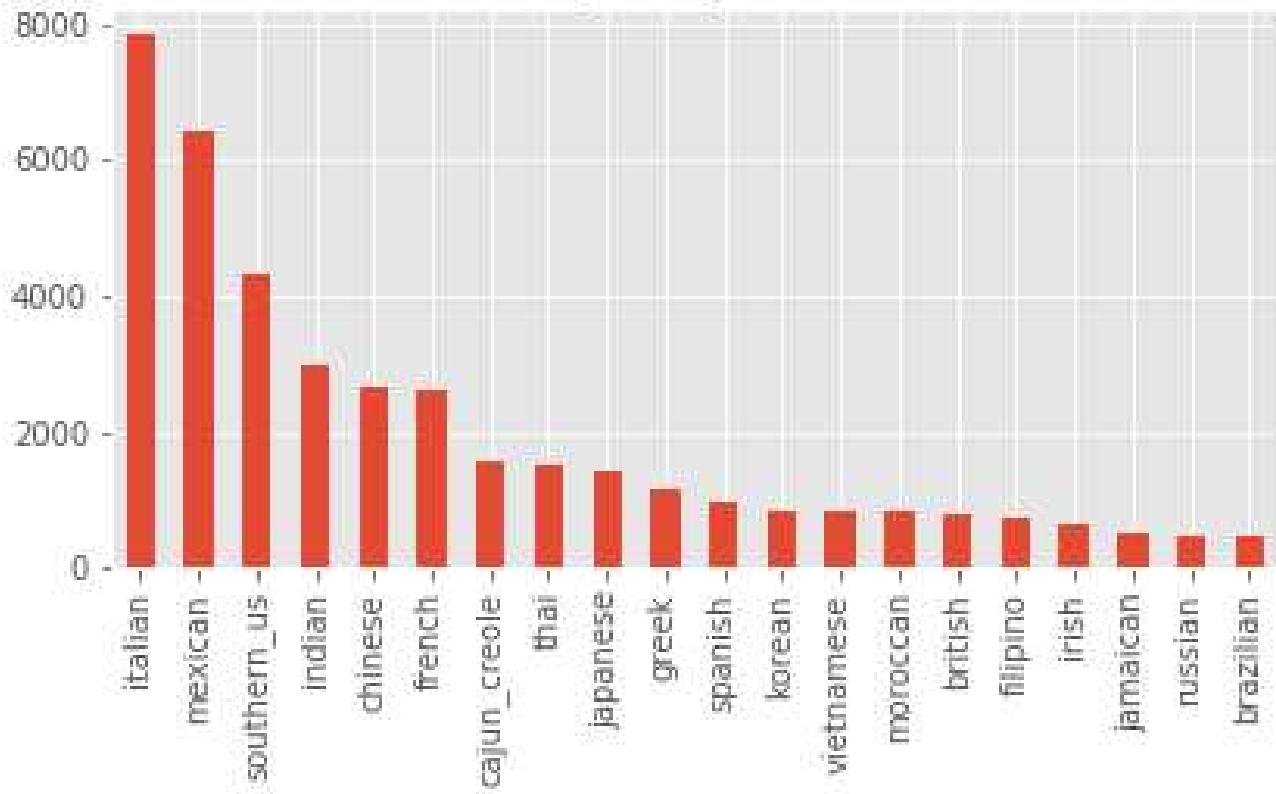
#read the ingredient data using pandas
dfTrain = pd.read_json('./data/train.json')

#load data using json
dataFilePath='./data/train.json'
with open(dataFilePath) as data_file:
    data = json.load(data_file)

```

**Figure 4: Data Loading**

### Recipies By Cuisine



**Figure 5: Recipe Distribution By Cuisine**

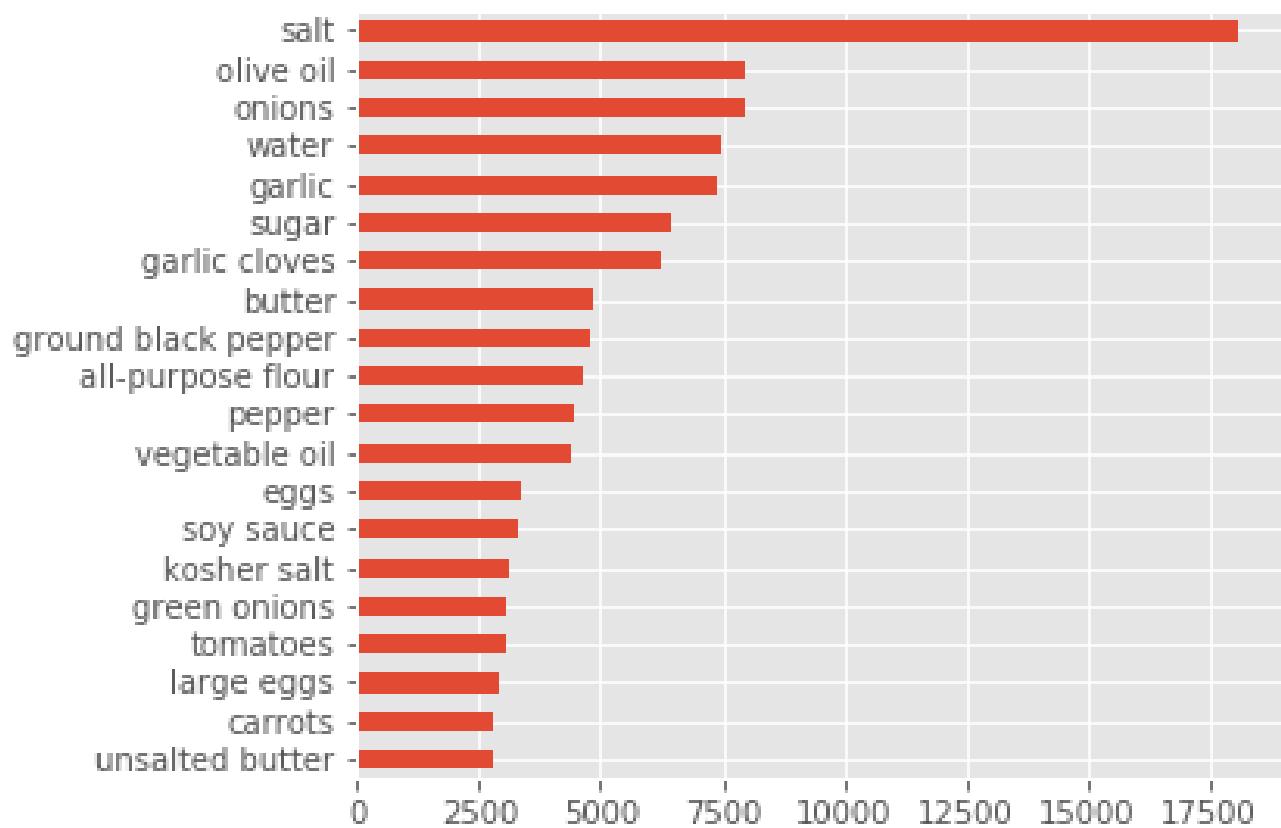


Figure 6: Top 20 Ingredients

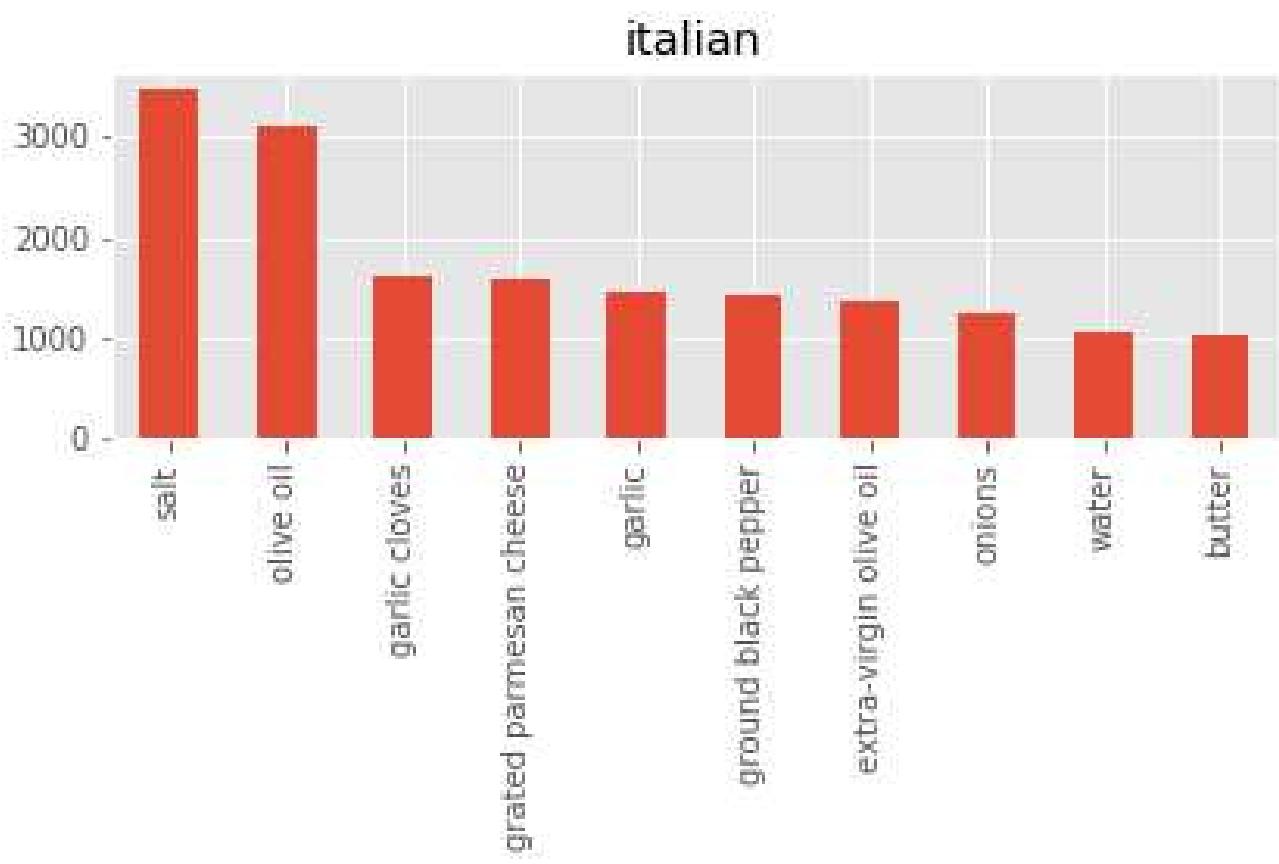


Figure 7: Top 10 Ingredients

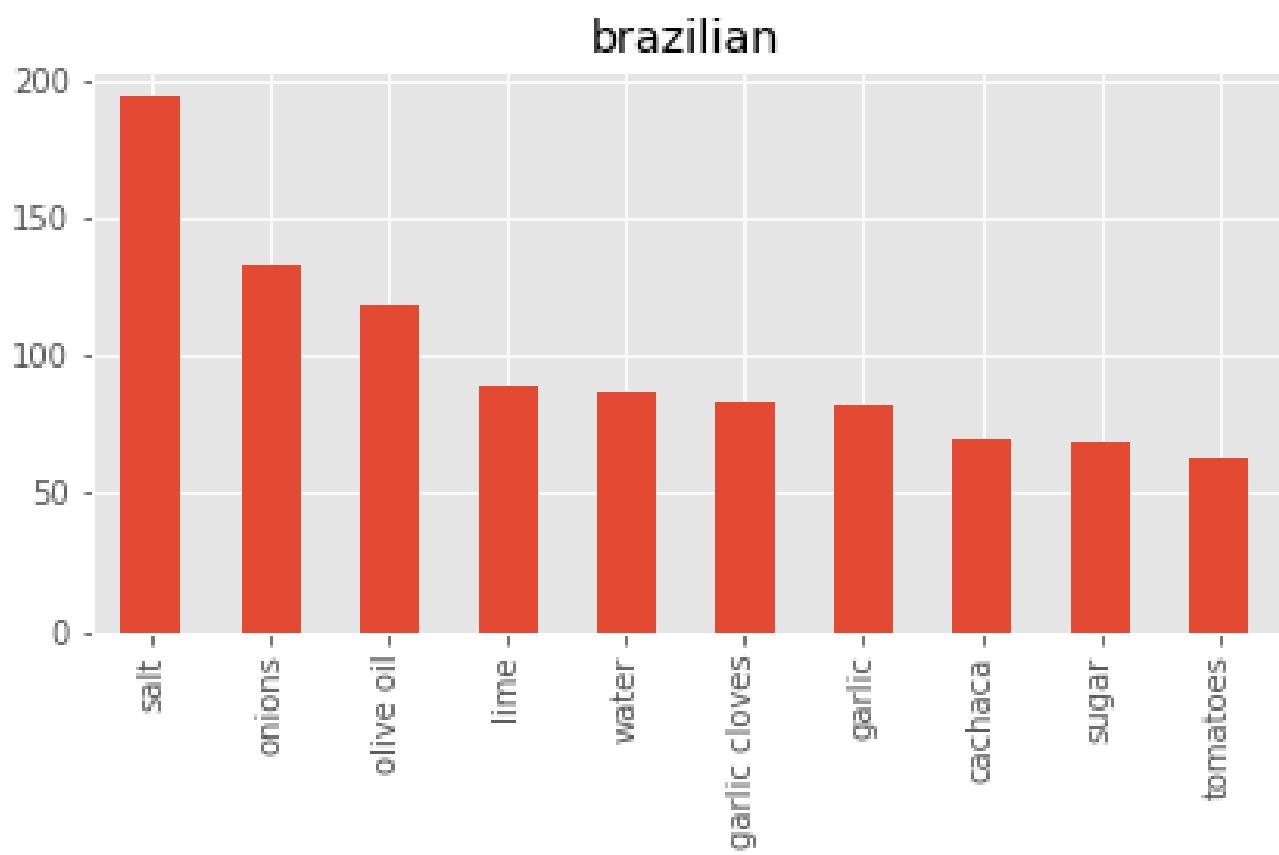
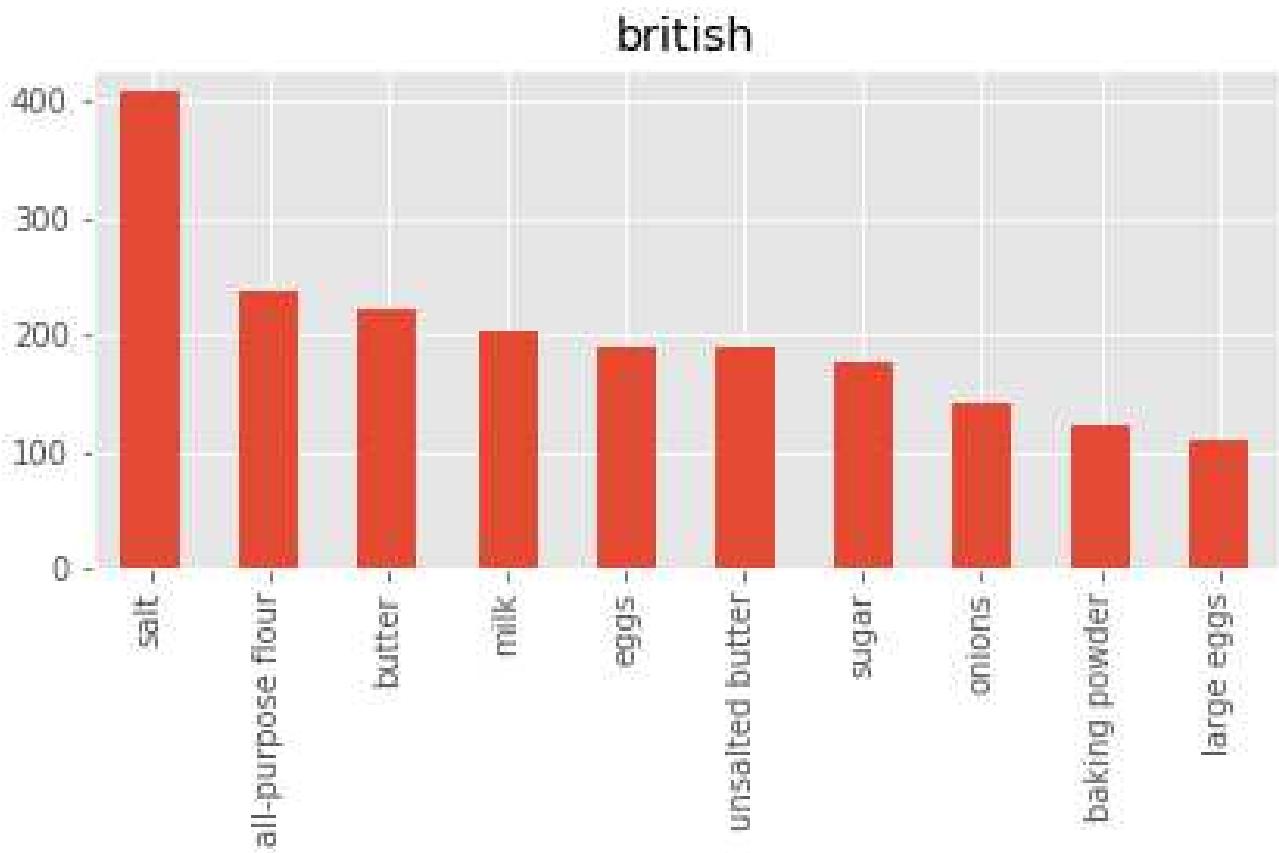


Figure 8: Top 10 Ingredients



**Figure 9: Top 10 Ingredients**

### cajun\_creole

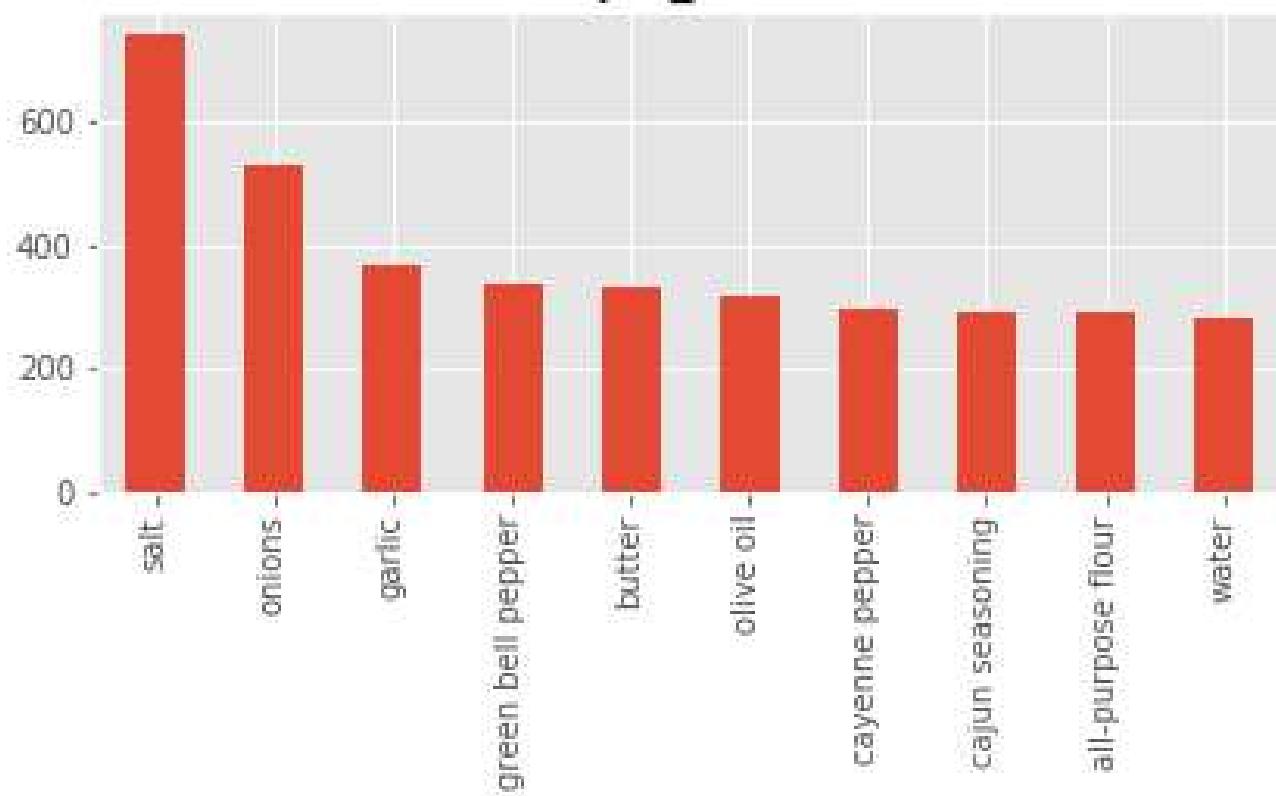


Figure 10: Top 10 Ingredients

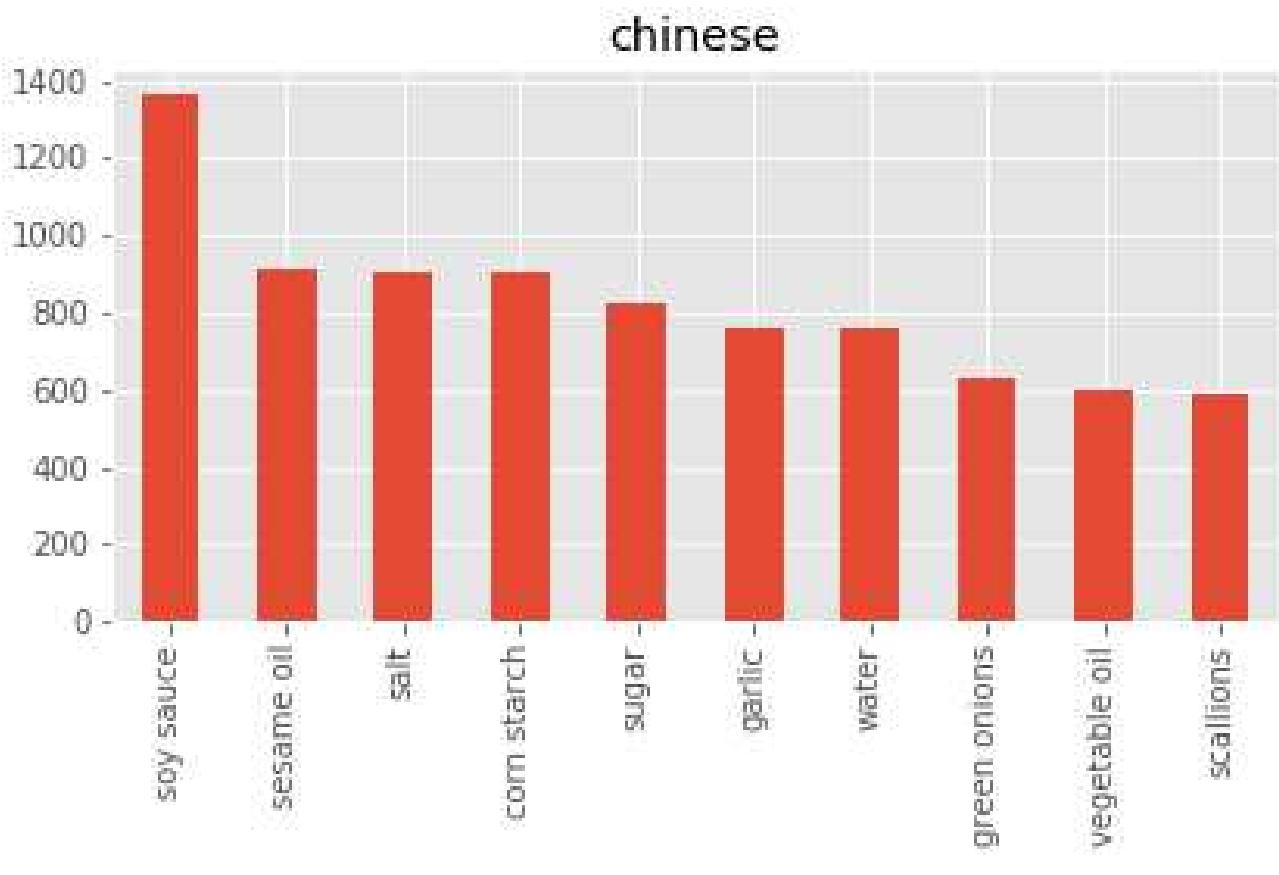


Figure 11: Top 10 Ingredients

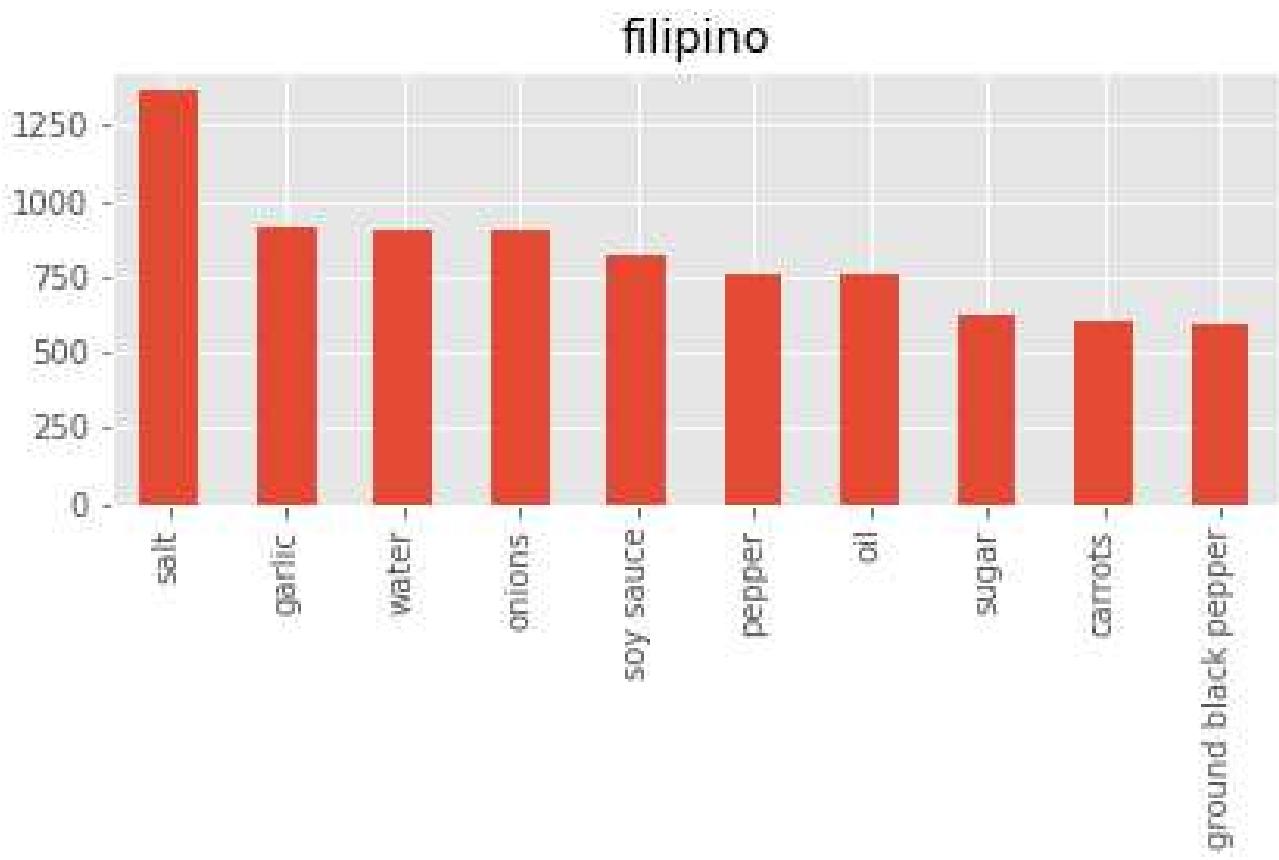


Figure 12: Top 10 Ingredients

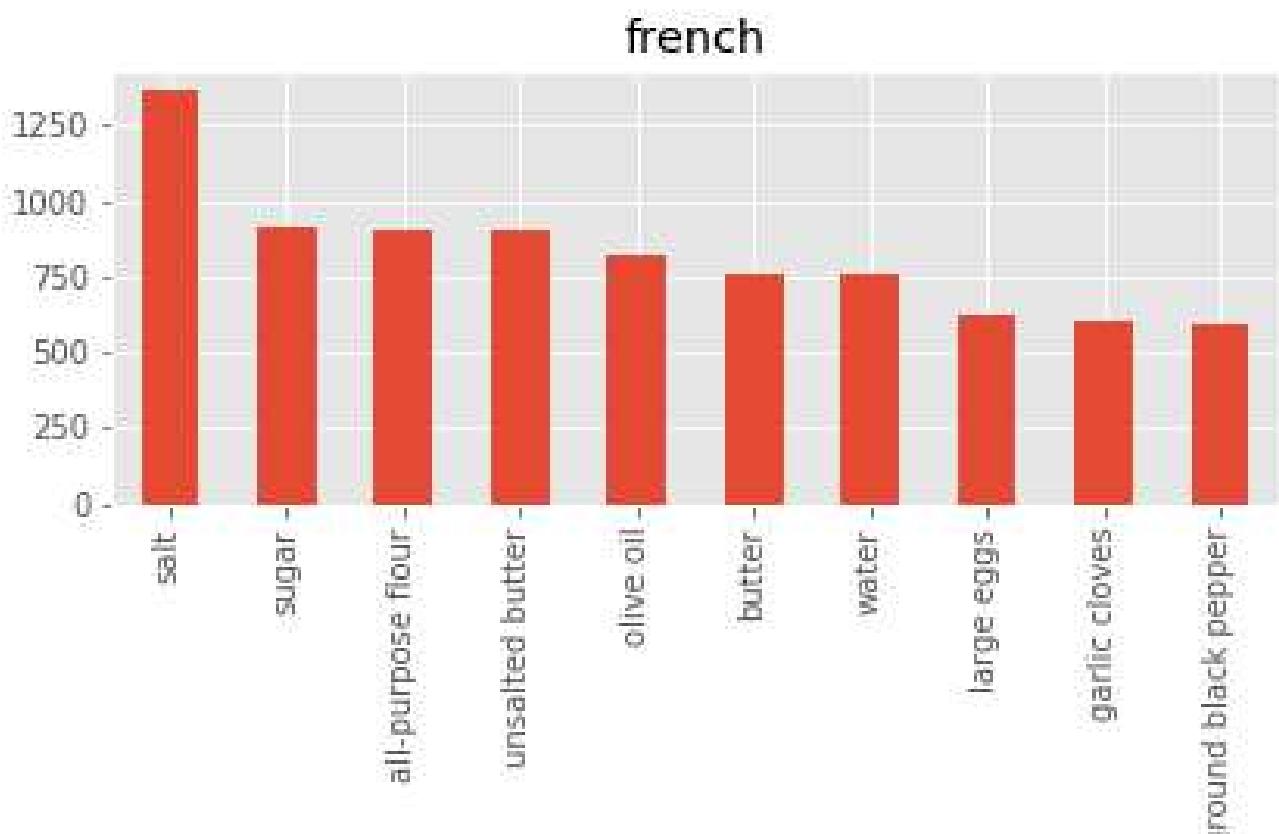


Figure 13: Top 10 Ingredients

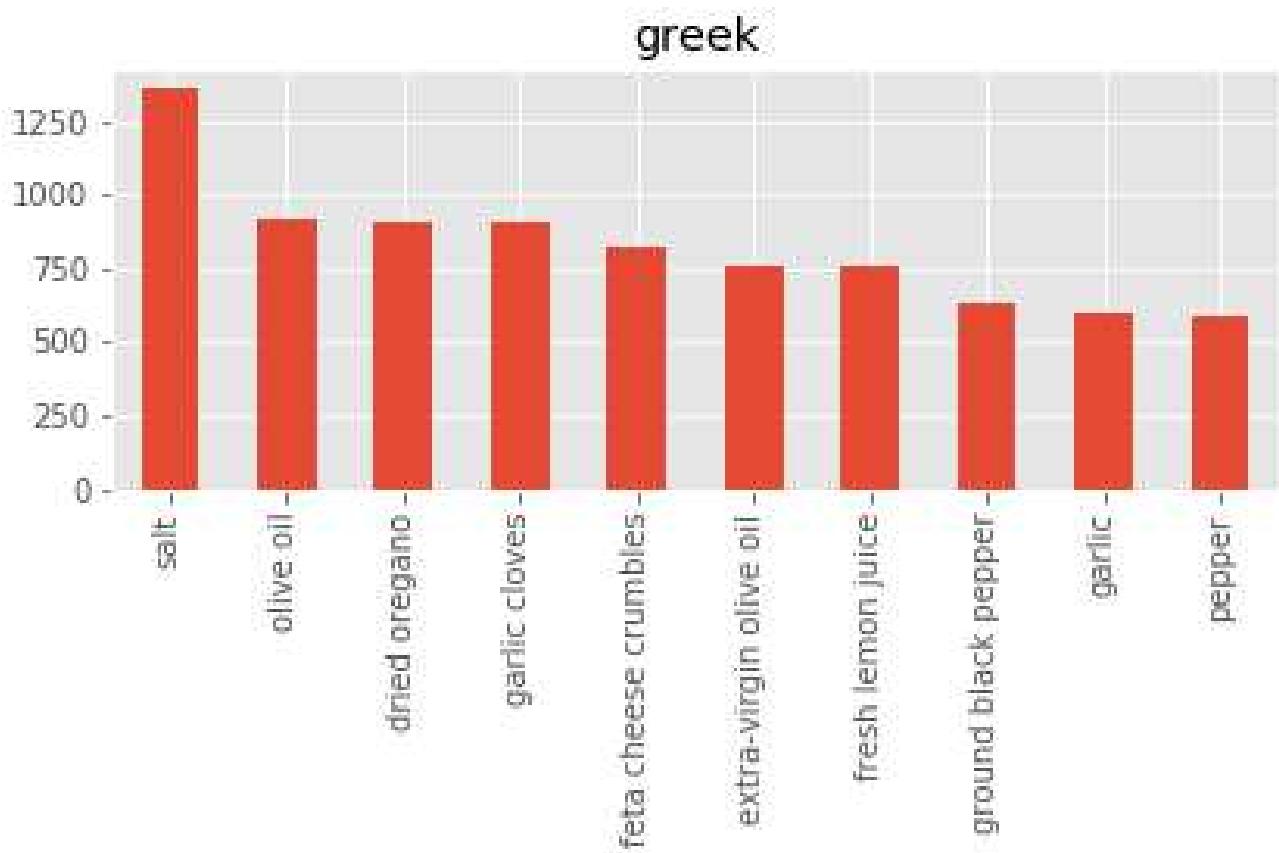
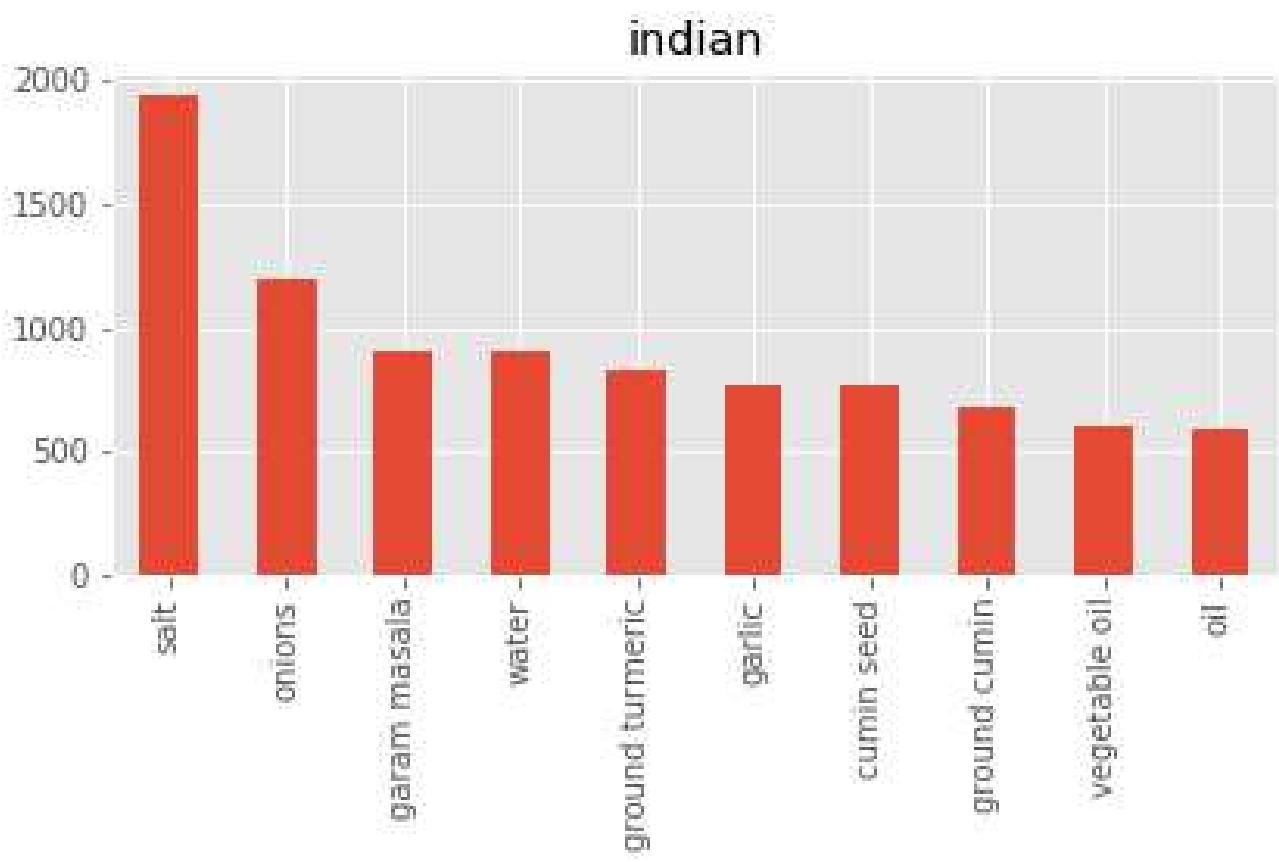


Figure 14: Top 10 Ingredients



**Figure 15: Top 10 Ingredients**

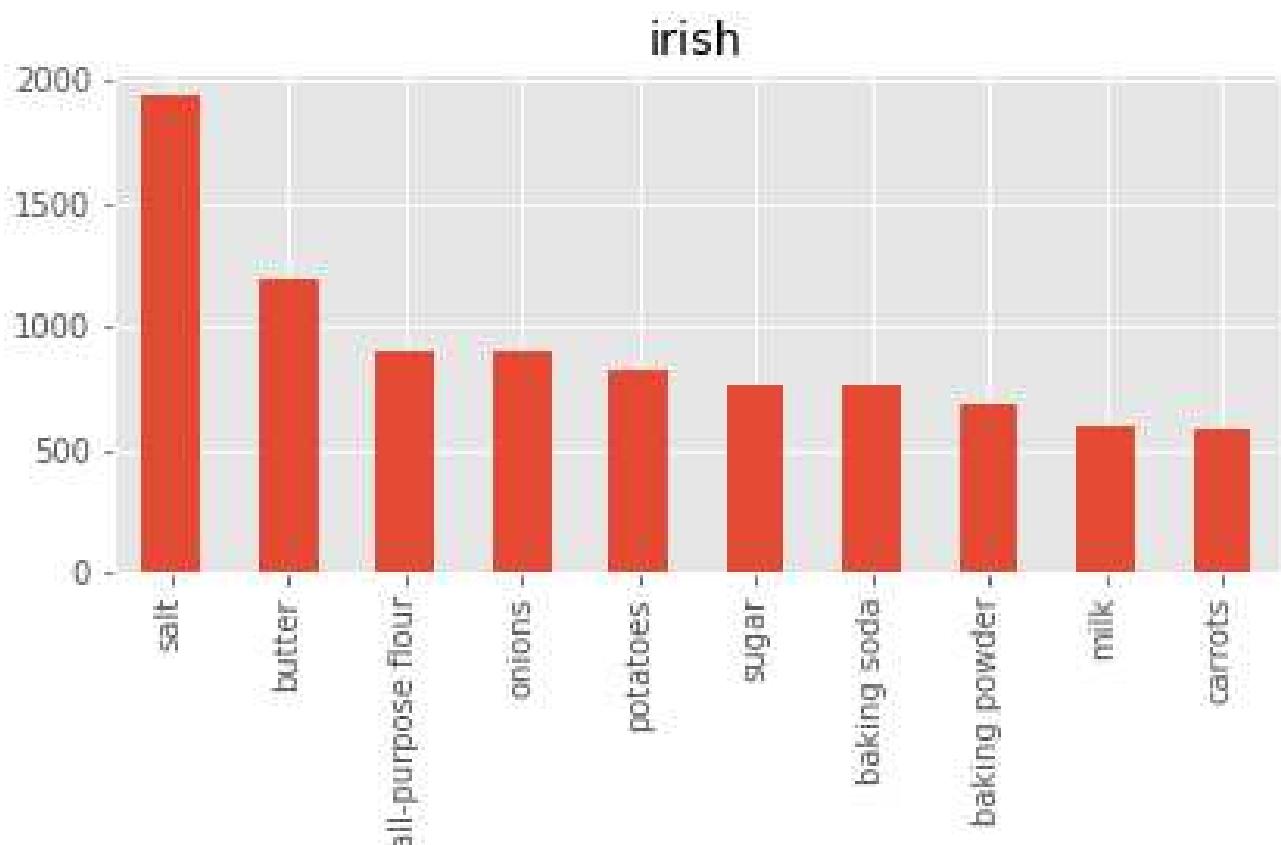


Figure 16: Top 10 Ingredients

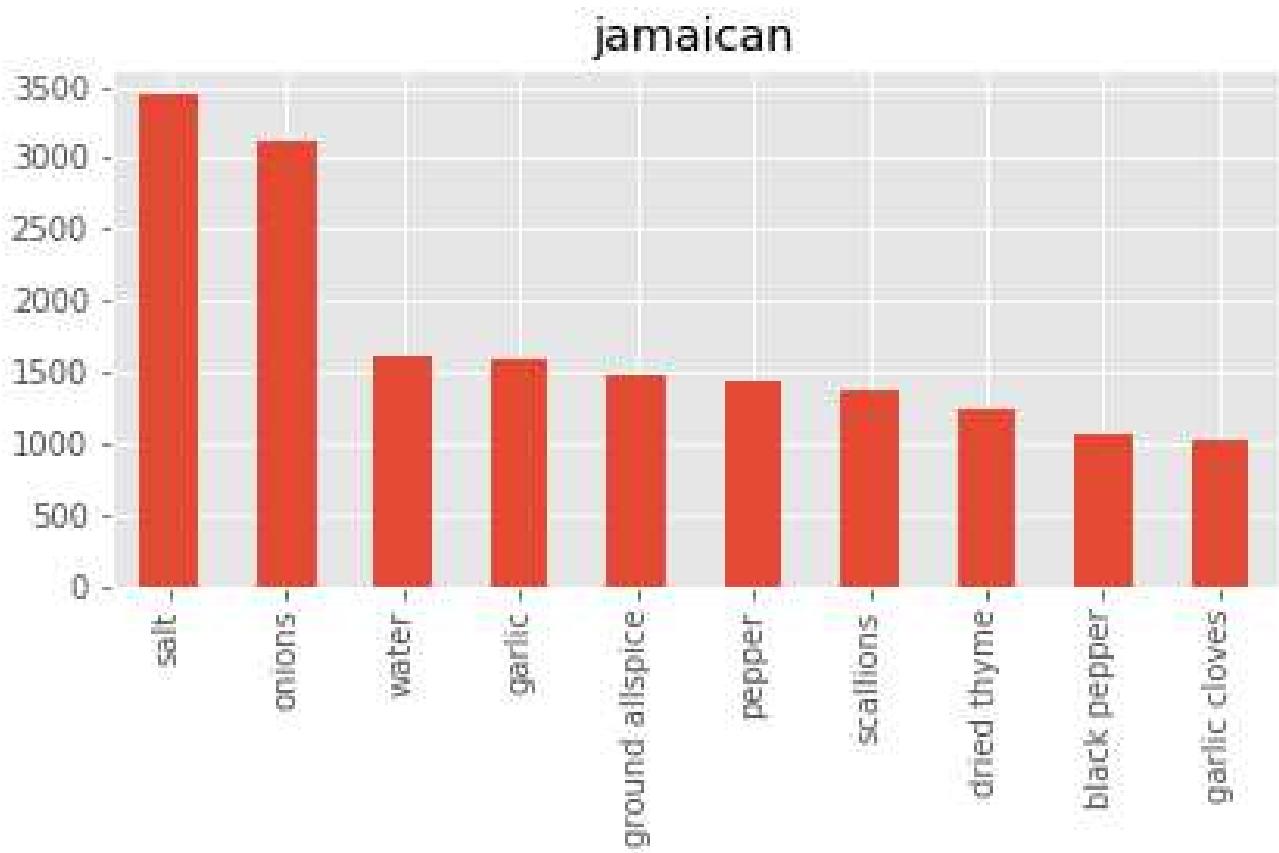


Figure 17: Top 10 Ingredients

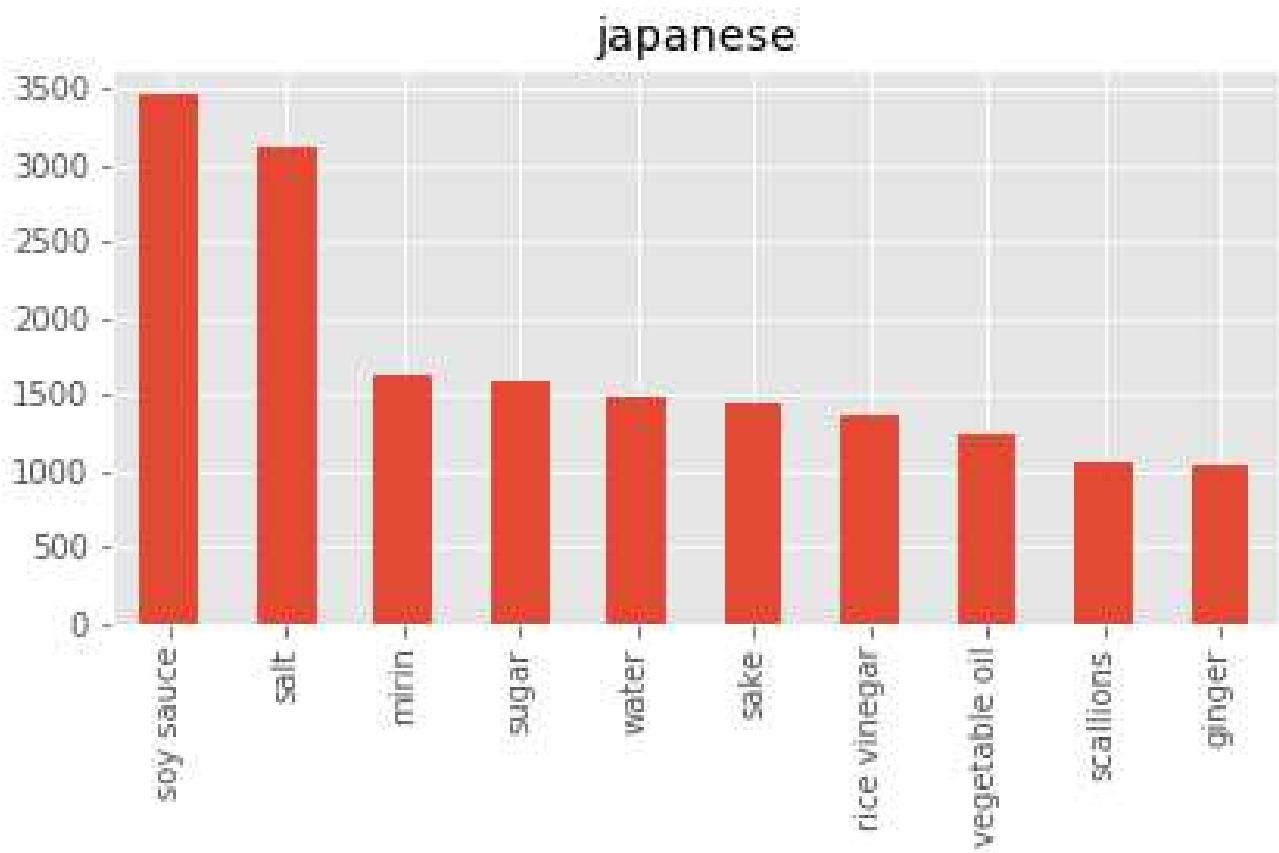


Figure 18: Top 10 Ingredients

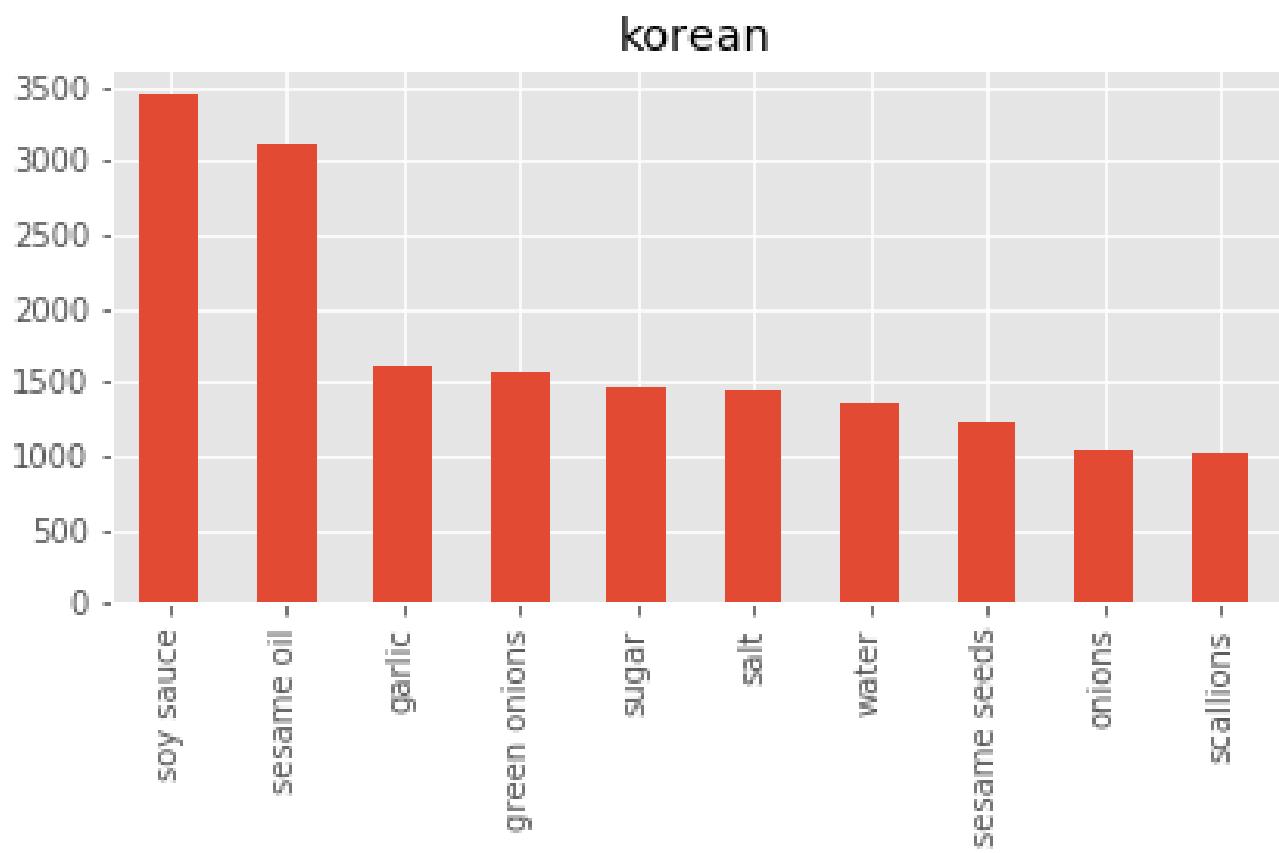


Figure 19: Top 10 Ingredients

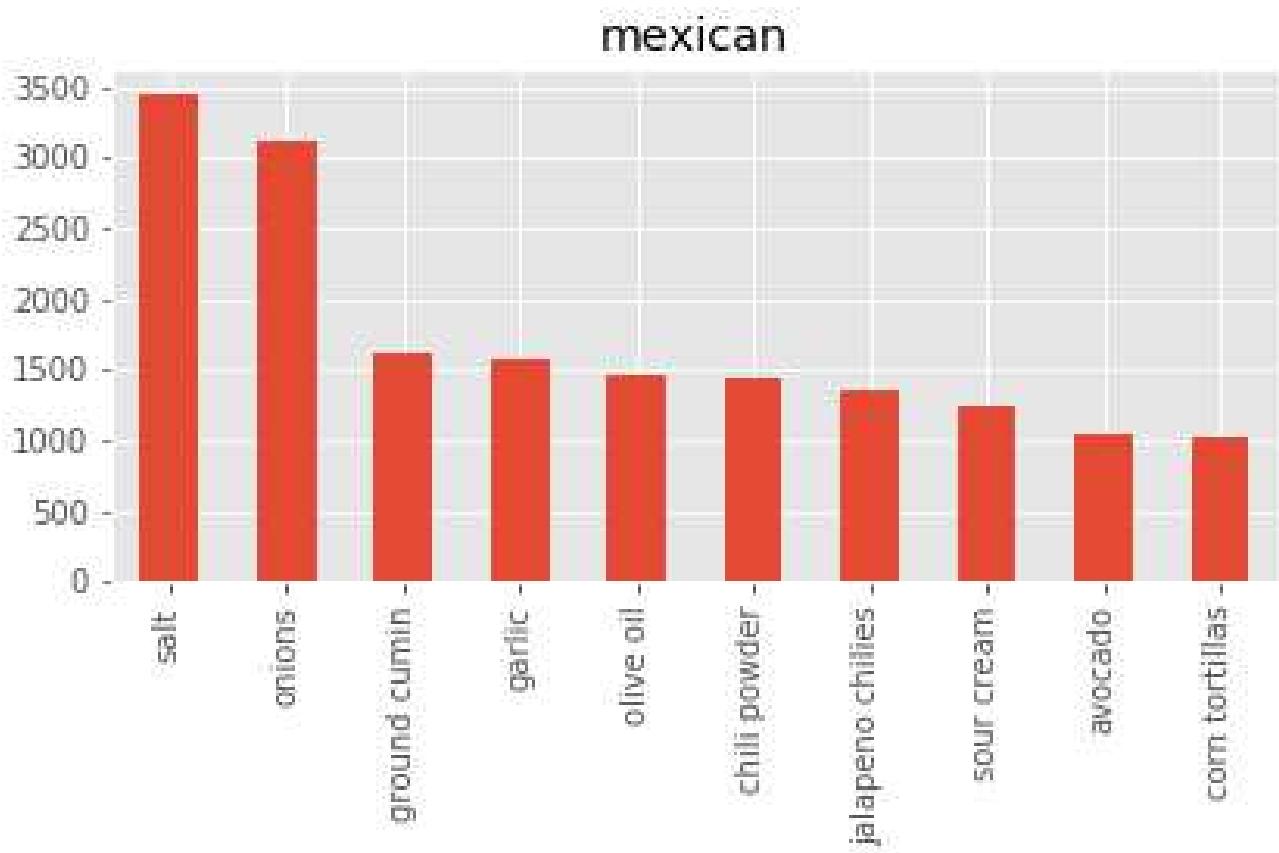


Figure 20: Top 10 Ingredients

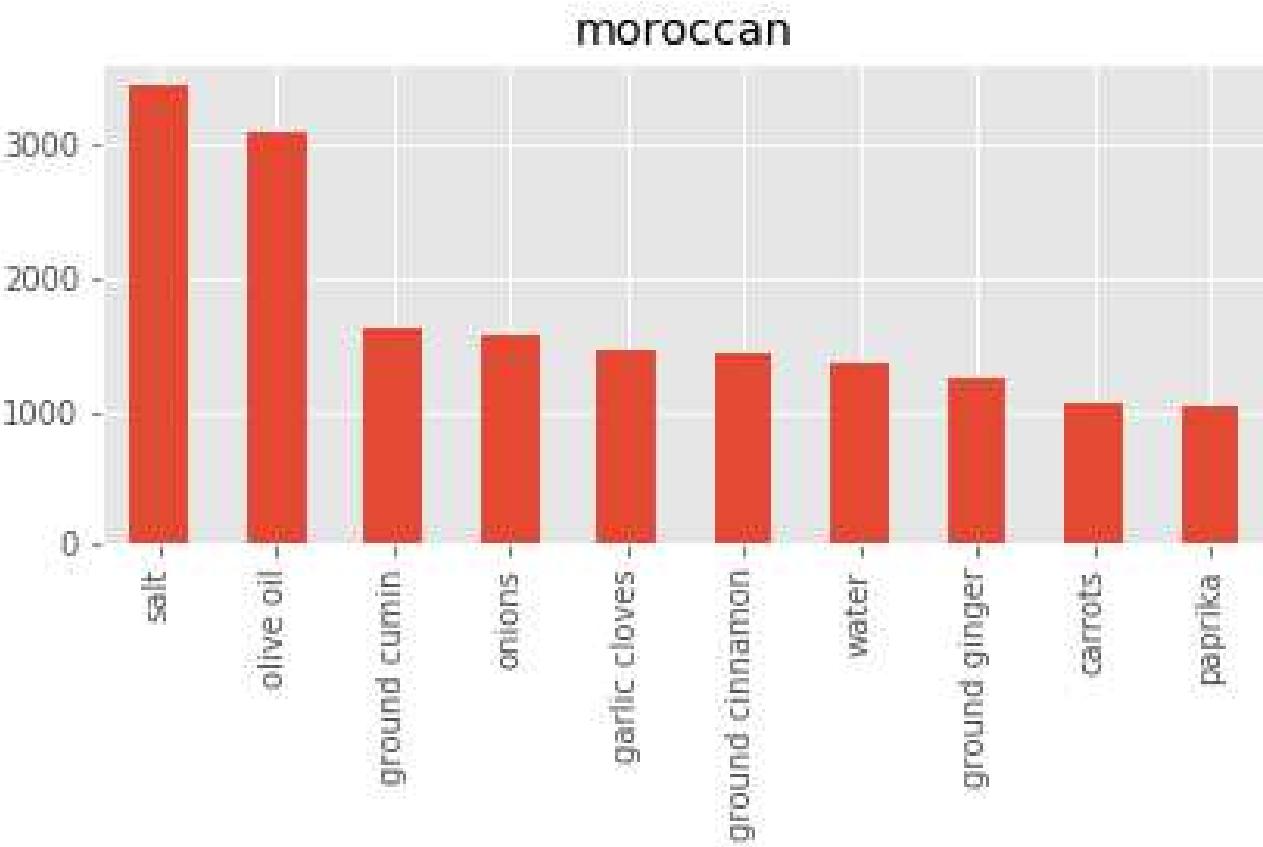


Figure 21: Top 10 Ingredients

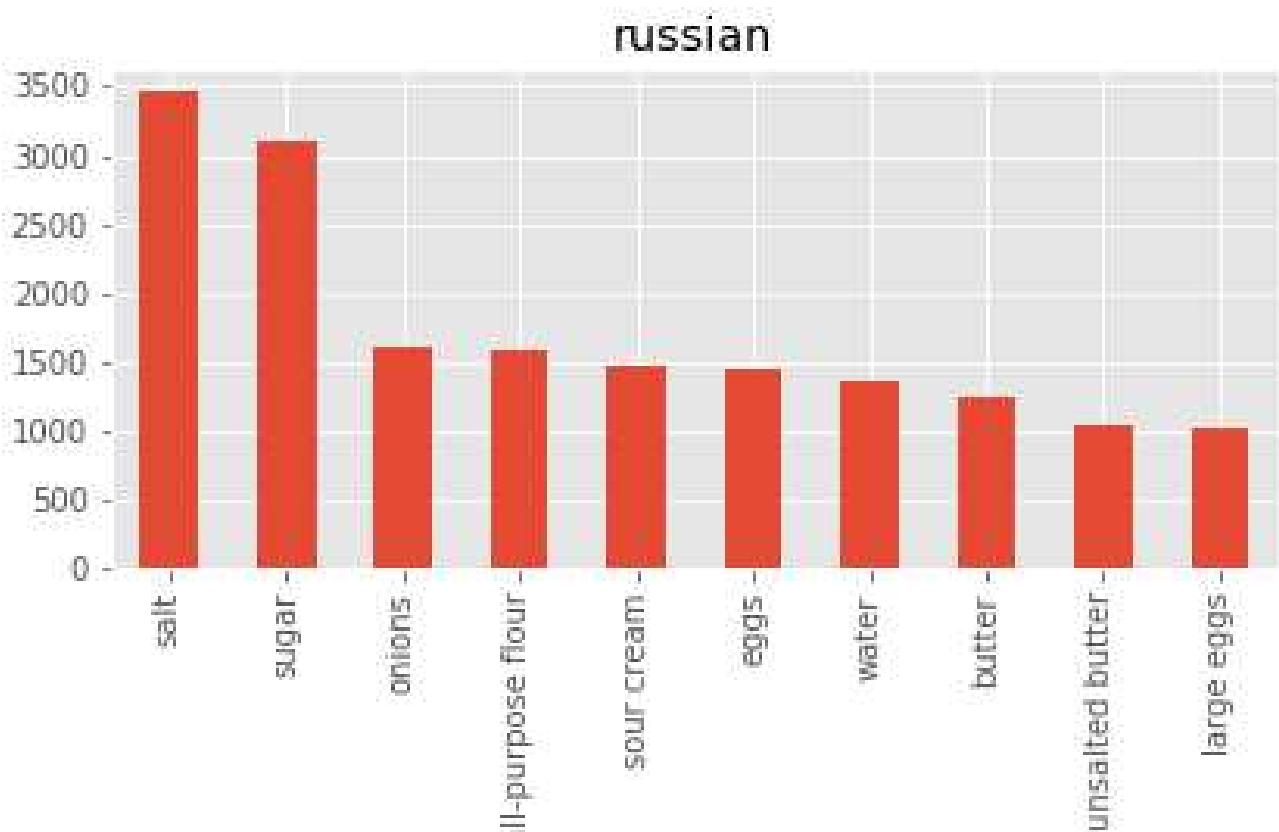


Figure 22: Top 10 Ingredients

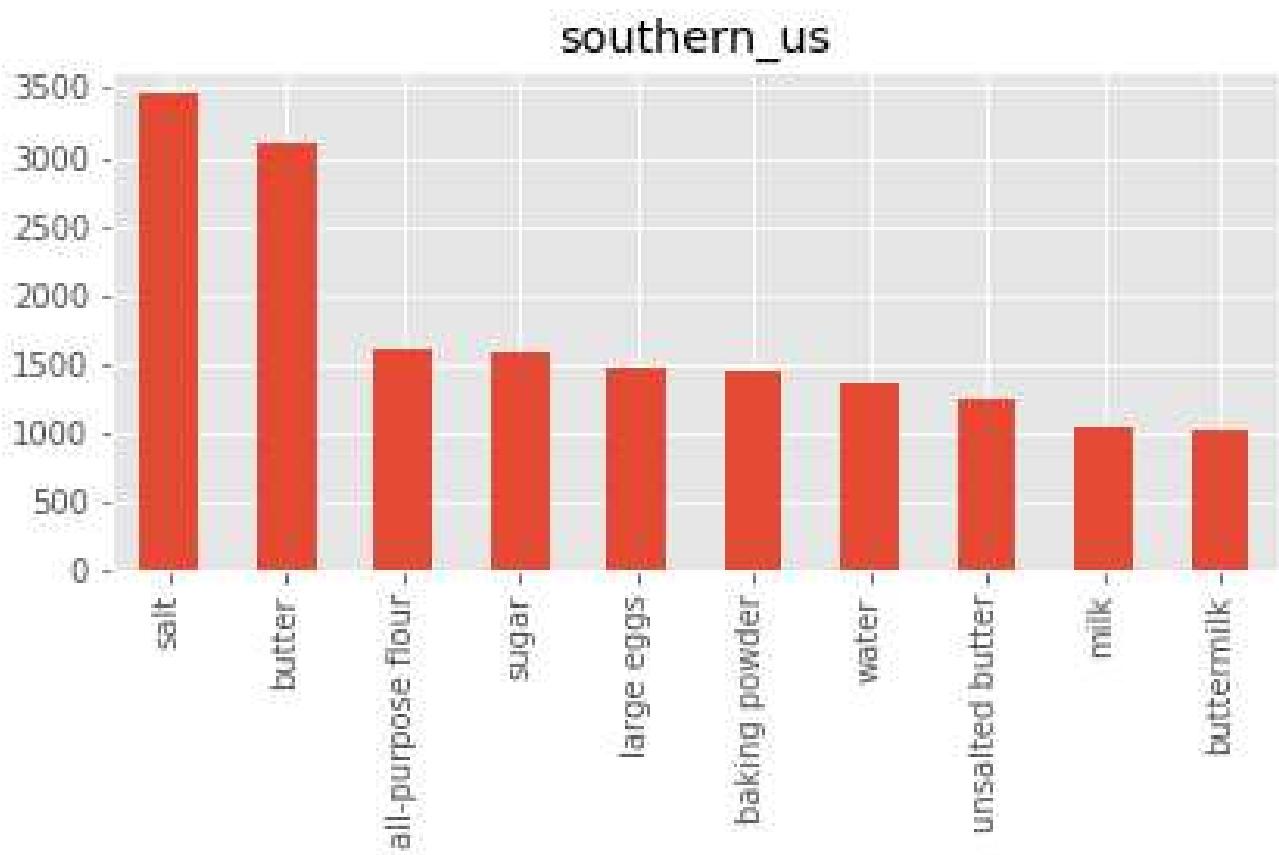


Figure 23: Top 10 Ingredients

spanish

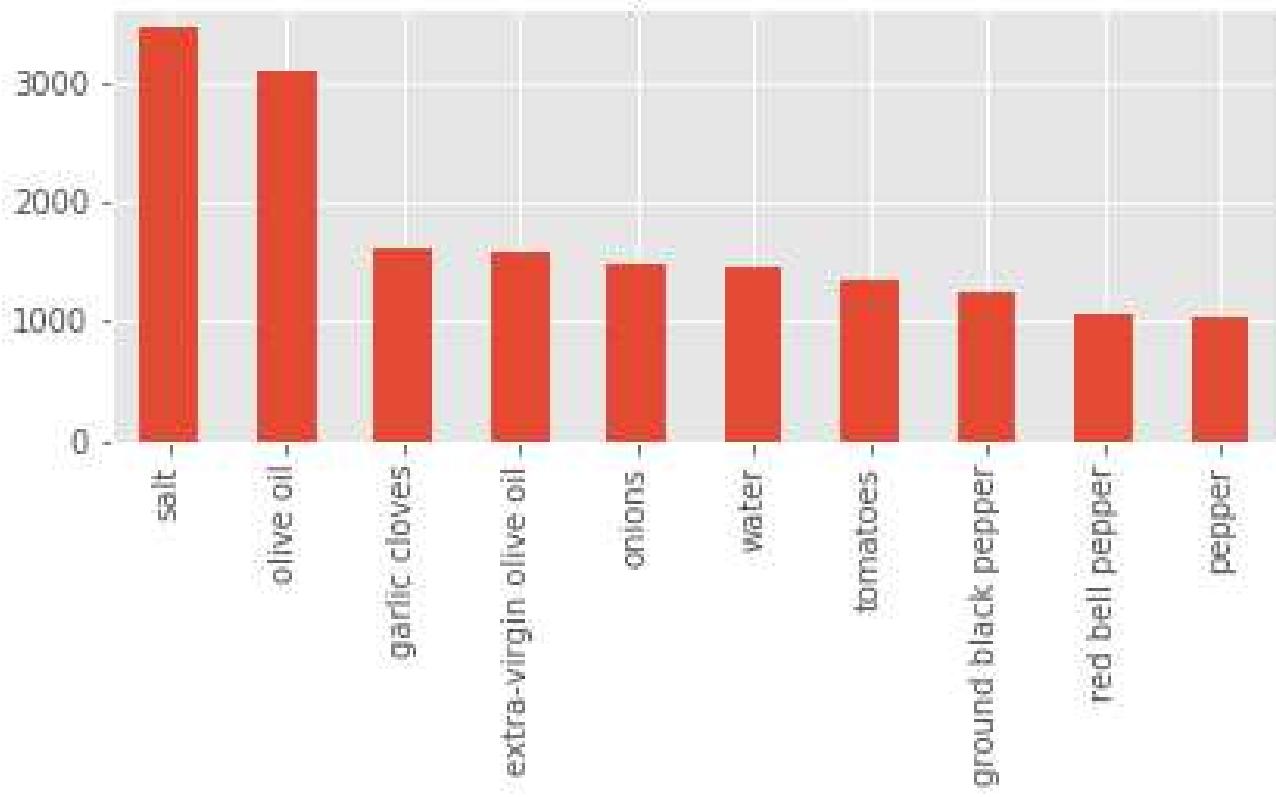


Figure 24: Top 10 Ingredients

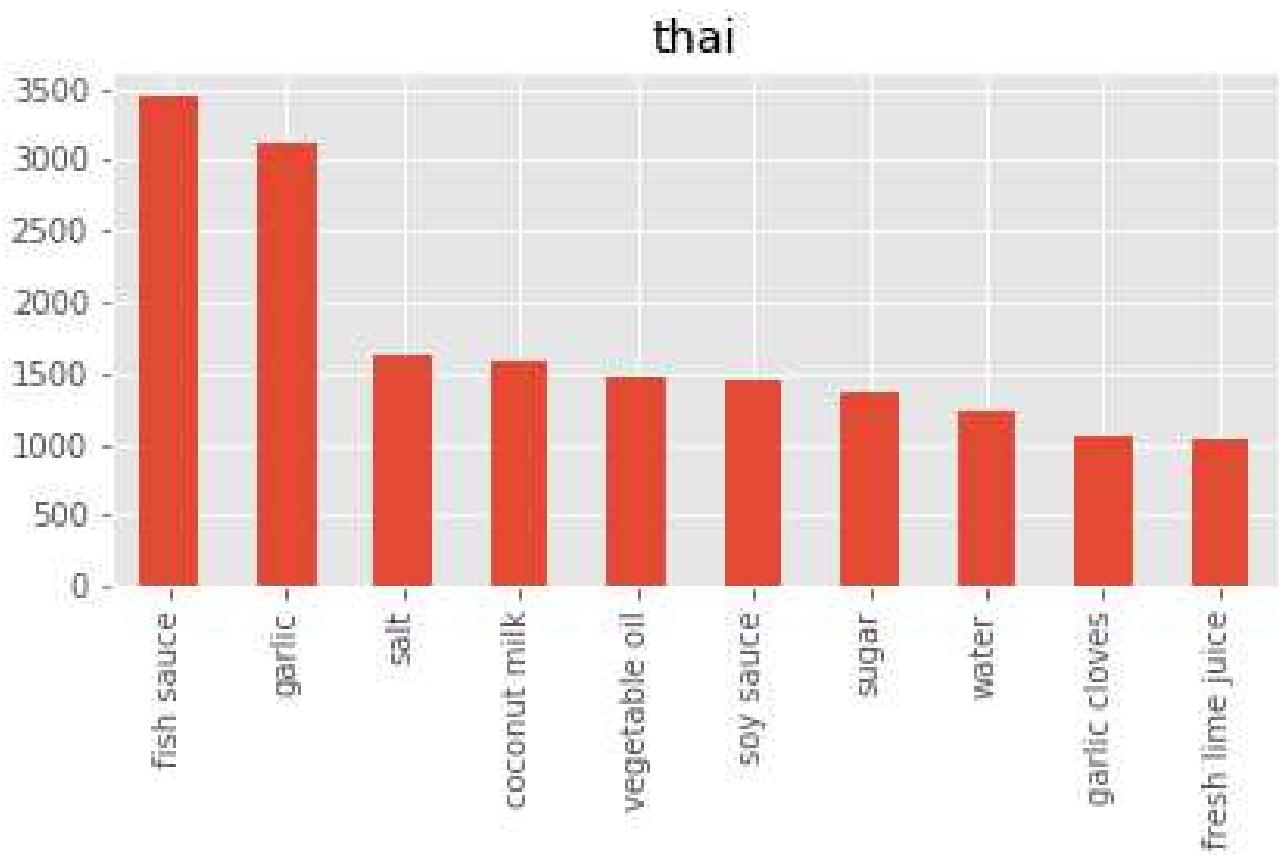


Figure 25: Top 10 Ingredients

## vietnamese

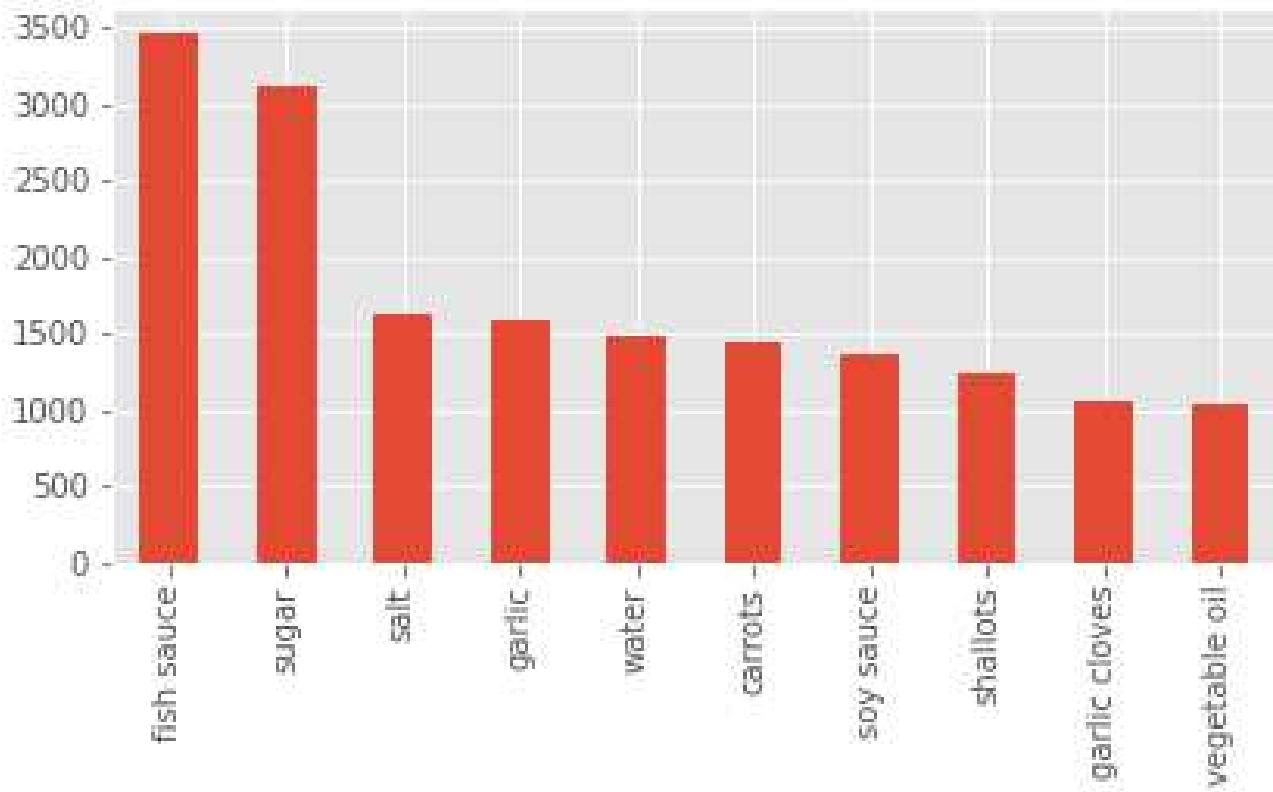
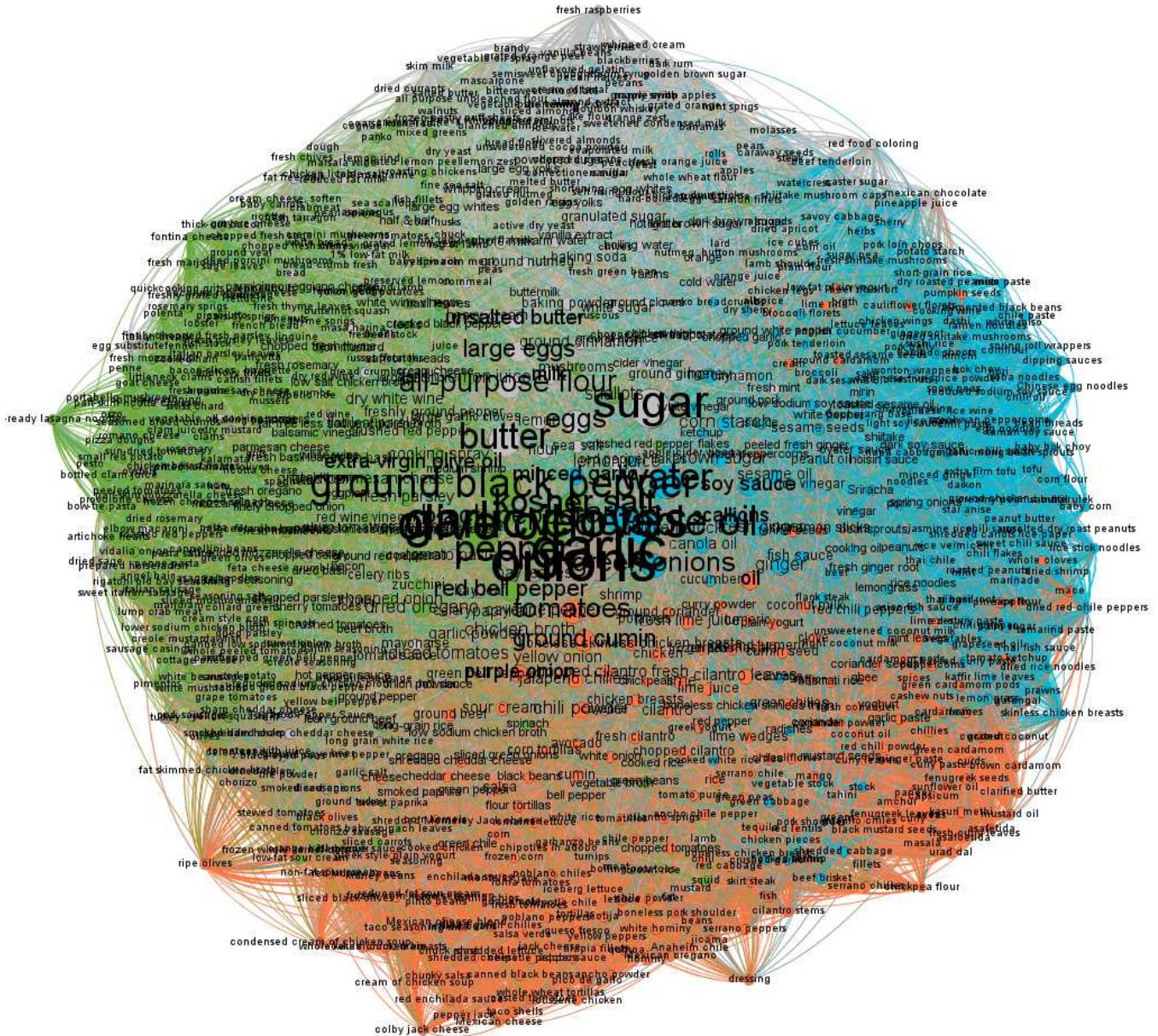
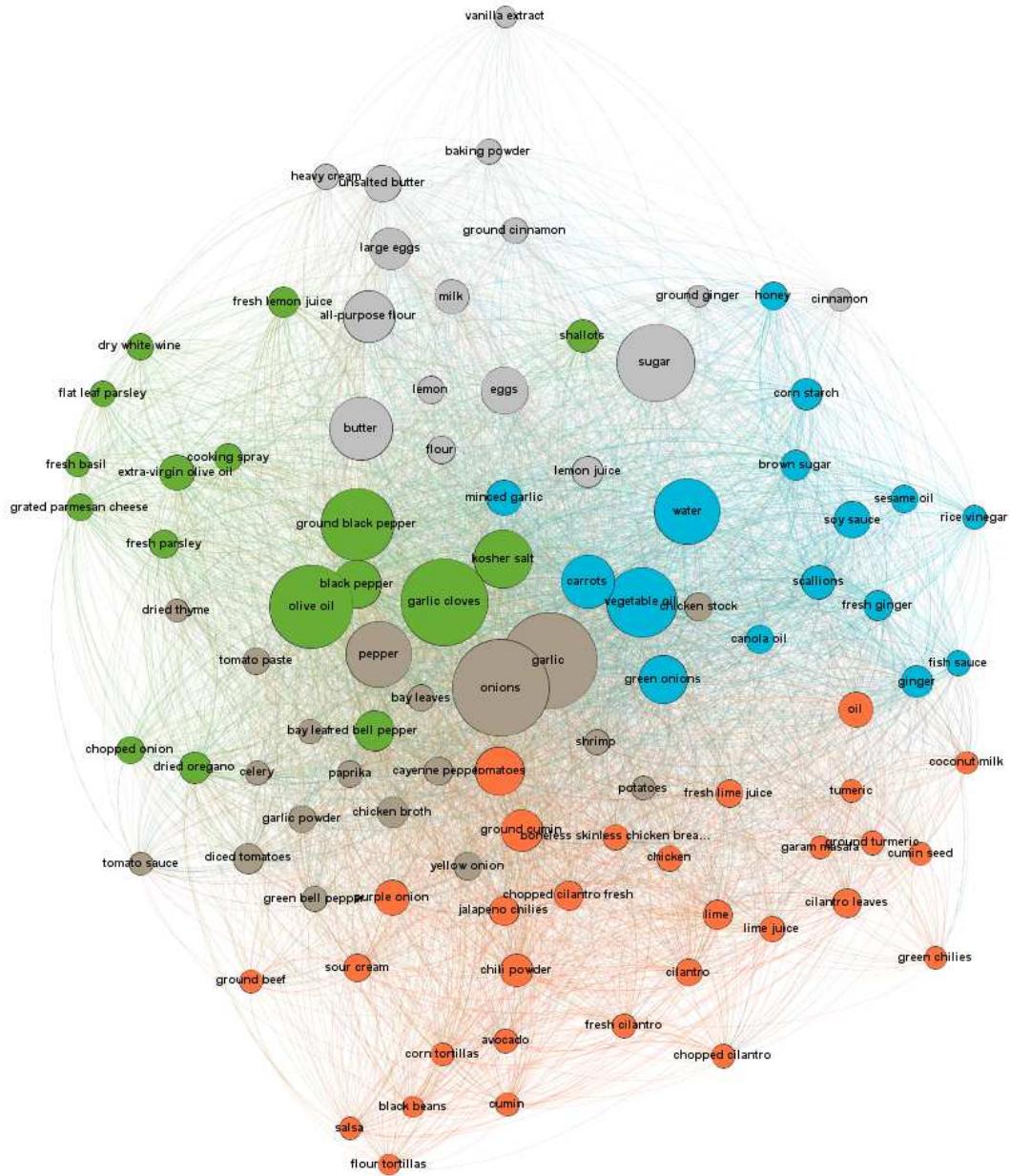


Figure 26: Top 10 Ingredients



**Figure 27: Ingredient Cluster**



**Figure 28: ingredient Cluster 100 Nodes**

LIST OF TABLES

1 Recipe Count By Cuisine

35

**Table 1: Recipe Count By Cuisine**

Cuisine	Recipe Count
brazilian	467
british	804
cajun creole	1546
chinese	2673
filipino	755
french	2646
greek	1175
indian	3003
irish	667
italian	7838
jamaican	526
japanese	1423
korean	830
mexican	6438
moroccan	821
russian	489
southern us	4320
spanish	989
thai	1539
vietnamese	825

```
bibtext report
```

```
=====
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
[2017-12-16 09.35.27] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.3s.
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Sushant Athaley
hid: 302
paper1: Nov 3 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

```
-----
```

wordcount

---

```
(null)
wc 302 project (null) 5586 report.tex
wc 302 project (null) 5781 report.pdf
wc 302 project (null) 573 report.bib
```

find "

---

```
163: "id": 24717,
164: "cuisine": "indian",
165: "ingredients": [
166:   "tumeric",
167:   "vegetable stock",
168:   "tomatoes",
169:   "garam masala",
170:   "naan",
171:   "red lentils",
172:   "red chili peppers",
173:   "onions",
174:   "spinach",
175:   "sweet potatoes"
189: dataFilePath="./data/train.json"
```

passed: False

find footnote

---

```
passed: True

find input{format/i523}
-----
passed: False

find input{format/final}
-----
passed: False

floats
-----
104: Code is organized as described in Figure \ref{c:code-structure}
105: \begin{figure}[htb]
118: \caption{Code Structure}\label{c:code-structure}
145: Figure \ref{f:methodology} shows pictorial representation of the
     methodology used for this project to analyze ingredient data.
146: \begin{figure}![ht]
147: \centering\includegraphics[width=\columnwidth]{images/methodology
     .PNG}
148: \caption{Flowchart of the Methodology to Analyze Ingredients
     }\label{f:methodology}
159: The dataset for this study is sourced from Kaggle application
     \cite{www-kaggle}. This dataset is publicly available and
     featured in \emph{What's Cooking?} competition. This dataset is
     provided to Kaggle by Yummly which is the application which hosts
     recipes online. This dataset is in JSON format and of 12MB size.
     This dataset contains recipe id, cuisine and list of ingredients
     as described in Figure \ref{c:data-structure}.
160: \begin{figure}[htb]
179: \caption{Ingredient Data Structure}\label{c:data-structure}
181: This dataset contains total 39774 recipes across various
     cuisines. We used two different methods to load this data.
     Cuisine and ingredient analysis is done by loading data into
     \emph{pandas dataframe} and to analyze ingredient relationship
     data has been loaded into \emph{json} object. Figure \ref{c:loading}
     shows the code for data loading used in this project.
182: \begin{figure}[htb]
193: \caption{Data Loading}\label{c:loading}
213: We first analyze entire dataset to understand the total number of
     recipes and their distribution across various cuisines. We use
     Python's Panda library to get the total recipe count as 39774 and
     plot the distribution. Figure
```

`\ref{f:Number_of_recipes_by_cuisine}` shows number of recipes per cuisine. Our observations from this analysis are  
 221: `\begin{figure}[!ht]`  
 222: `\centering\includegraphics[width=\columnwidth]{images/Number_of_recipes_by_cuisine.png}`  
 223: `\caption{Recipe Distribution By Cuisine}\label{f:Number_of_recipes_by_cuisine}`  
 226: Table `\ref{t:recipecount}` describes recipe count for every cuisine.  
 227: `\begin{table}[htb]`  
 230: `\label{t:recipecount}`  
 278: Ingredient `\emph{Salt}` is obvious topper followed by `\emph{Oil}` and `\emph{Onions}`. This also proves our craving for salty and fatty food. Top 20 ingredient also contain duplicate ingredient like garlic and garlic clove, salt and kosher salt, eggs and large eggs which shows shortcoming of the dataset. Also ingredient like salt, oil and water could be avoided to get analysis of real ingredients as these are commonly used ingredient and doesn't contribute much to the study. Figure `\ref{f:Ingredient_Distribution}` shows top 20 ingredient across cuisines.  
 279: `\begin{figure}[!ht]`  
 280: `\centering\includegraphics[width=\columnwidth]{images/Ingredient_Distribution.png}`  
 281: `\caption{Top 20 Ingredients }\label{f:Ingredient_Distribution}`  
 294: Similarly we show key ingredient of all other cuisines present in the dataset and we observe that it is very close representation of all cuisines. Figure `\ref{f:italian_10_most_used_ingredients}`, `\ref{f:brazilian_10_most_used_ingredients}`, `\ref{f:british_10_most_used_ingredients}`, `\ref{f:cajun_creole_10_most_used_ingredients}`, `\ref{f:chinese_10_most_used_ingredients}`, `\ref{f:filipino_10_most_used_ingredients}`, `\ref{f:french_10_most_used_ingredients}`, `\ref{f:greek_10_most_used_ingredients}`, `\ref{f:indian_10_most_used_ingredients}`, `\ref{f:irish_10_most_used_ingredients}`, `\ref{f:jamaican_10_most_used_ingredients}`, `\ref{f:japanese_10_most_used_ingredients}`, `\ref{f:korean_10_most_used_ingredients}`, `\ref{f:mexican_10_most_used_ingredients}`, `\ref{f:moroccan_10_most_used_ingredients}`, `\ref{f:russian_10_most_used_ingredients}`, `\ref{f:southern_us_10_most_used_ingredients}`, `\ref{f:spanish_10_most_used_ingredients}`, `\ref{f:thai_10_most_used_ingredients}`,

```

\ref{f:vietnamese_10_most_used_ingredients} shows top 10 key
ingredient used in the corresponding cuisines.

295: \begin{figure}[!ht]
296: \centering\includegraphics[width=\columnwidth]{images/italian_10_
most_used_ingredients.png}
297: \caption{Top 10 Ingredients
} \label{f:italian_10_most_used_ingredients}
300: \begin{figure}[!ht]
301: \centering\includegraphics[width=\columnwidth]{images/brazilian_1
0_most_used_ingredients.png}
302: \caption{Top 10 Ingredients
} \label{f:brazilian_10_most_used_ingredients}
305: \begin{figure}[!ht]
306: \centering\includegraphics[width=\columnwidth]{images/british_10_
most_used_ingredients.png}
307: \caption{Top 10 Ingredients
} \label{f:british_10_most_used_ingredients}
310: \begin{figure}[!ht]
311: \centering\includegraphics[width=\columnwidth]{images/cajun_creol
e_10_most_used_ingredients.png}
312: \caption{Top 10 Ingredients
} \label{f:cajun_creole_10_most_used_ingredients}
315: \begin{figure}[!ht]
316: \centering\includegraphics[width=\columnwidth]{images/chinese_10_
most_used_ingredients.png}
317: \caption{Top 10 Ingredients
} \label{f:chinese_10_most_used_ingredients}
320: \begin{figure}[!ht]
321: \centering\includegraphics[width=\columnwidth]{images/filipino_10
_most_used_ingredients.png}
322: \caption{Top 10 Ingredients
} \label{f:filipino_10_most_used_ingredients}
325: \begin{figure}[!ht]
326: \centering\includegraphics[width=\columnwidth]{images/french_10_m
ost_used_ingredients.png}
327: \caption{Top 10 Ingredients
} \label{f:french_10_most_used_ingredients}
330: \begin{figure}[!ht]
331: \centering\includegraphics[width=\columnwidth]{images/greek_10_mo
st_used_ingredients.png}
332: \caption{Top 10 Ingredients
} \label{f:greek_10_most_used_ingredients}
335: \begin{figure}[!ht]
336: \centering\includegraphics[width=\columnwidth]{images/indian_10_m
ost_used_ingredients.png}
337: \caption{Top 10 Ingredients
}

```

```

} \label{f:indian_10_most_used_ingredients}
340: \begin{figure} [!ht]
341: \centering \includegraphics [width=\columnwidth] {images/irish_10_mo
st_used_ingredients.png}
342: \caption{Top 10 Ingredients
} \label{f:irish_10_most_used_ingredients}
345: \begin{figure} [!ht]
346: \centering \includegraphics [width=\columnwidth] {images/jamaican_10
_most_used_ingredients.png}
347: \caption{Top 10 Ingredients
} \label{f:jamaican_10_most_used_ingredients}
350: \begin{figure} [!ht]
351: \centering \includegraphics [width=\columnwidth] {images/japanese_10
_most_used_ingredients.png}
352: \caption{Top 10 Ingredients
} \label{f:japanese_10_most_used_ingredients}
355: \begin{figure} [!ht]
356: \centering \includegraphics [width=\columnwidth] {images/korean_10_m
ost_used_ingredients.png}
357: \caption{Top 10 Ingredients
} \label{f:korean_10_most_used_ingredients}
360: \begin{figure} [!ht]
361: \centering \includegraphics [width=\columnwidth] {images/mexican_10_
most_used_ingredients.png}
362: \caption{Top 10 Ingredients
} \label{f:mexican_10_most_used_ingredients}
365: \begin{figure} [!ht]
366: \centering \includegraphics [width=\columnwidth] {images/moroccan_10
_most_used_ingredients.png}
367: \caption{Top 10 Ingredients
} \label{f:moroccan_10_most_used_ingredients}
370: \begin{figure} [!ht]
371: \centering \includegraphics [width=\columnwidth] {images/russian_10_
most_used_ingredients.png}
372: \caption{Top 10 Ingredients
} \label{f:russian_10_most_used_ingredients}
375: \begin{figure} [!ht]
376: \centering \includegraphics [width=\columnwidth] {images/southern_us
_10_most_used_ingredients.png}
377: \caption{Top 10 Ingredients
} \label{f:southern_us_10_most_used_ingredients}
380: \begin{figure} [!ht]
381: \centering \includegraphics [width=\columnwidth] {images/spanish_10_
most_used_ingredients.png}
382: \caption{Top 10 Ingredients
} \label{f:spanish_10_most_used_ingredients}

```

```

385: \begin{figure} [!ht]
386: \centering\includegraphics[width=\columnwidth]{images/thai_10_mos
t_used_ingredients.png}
387: \caption{Top 10 Ingredients
}\label{f:thai_10_most_used_ingredients}
390: \begin{figure} [!ht]
391: \centering\includegraphics[width=\columnwidth]{images/vietnamese_
10_most_used_ingredients.png}
392: \caption{Top 10 Ingredients
}\label{f:vietnamese_10_most_used_ingredients}
421: Figure \ref{f:ingredient_modularity} shows ingredient cluster of
more than 1000 nodes. This graph is nice to look at but difficult
to read due to lot many nodes and edges in the graph.
422: \begin{figure} [!ht]
423: \centering\includegraphics[width=\columnwidth]{images/ingredient_
modularity.png}
424: \caption{Ingredient Cluster }\label{f:ingredient_modularity}
427: Figure \ref{f:ingredient_modularity100} shows ingredient cluster
of around 100 nodes. We generated this graph by reducing nodes
and edges to make it more readable. This graph provides us with
our top 5 cuisine clusters.
428: \begin{figure} [!ht]
429: \centering\includegraphics[width=\columnwidth]{images/ingredient_
modularity100.png}
430: \caption{ingredient Cluster 100 Nodes
}\label{f:ingredient_modularity100}

```

figures 28

tables 1

includegraphics 25

labels 29

refs 10

floats 29

True : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

False : check if all figures are refered to: (refs >= labels)

Label/ref check

passed: True

When using figures use columnwidth

[width=1.0\columnwidth]

do not cahnge the number to a smaller fraction

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Killings by Police in the United States

Jeramy Townsley  
Butler University  
4600 Sunset Ave  
Indianapolis, Indiana 46208  
jtownsle@butler.edu

## ABSTRACT

With the rise of camera phones that allow citizens to videotape law enforcement brutality against citizens, and the ability to immediately make those videos public through social media, there has been an increased awareness of extent of police killings in the United States. With this new data there have been a number of systematic attempts to collect a record of these events from journalists, academics, and activists, since there is no publicly available, credible and full government database documenting these events. These deaths can be mapped at the county level with Quantum GIS, and demographic and economic data obtained from the Census at the county level can be used to create a regression model of these events with R. Evidence shows that structural economic factors are significant predictor variables to county-level killings by police.

## KEYWORDS

i523, hid347, police violence, negative binomial regression, gis mapping, American Community Survey

## 1 INTRODUCTION

In an era of seemingly ubiquitous tracking of human behavior by overlapping layers of government, one would expect there would be reliable data on homicides. Further, one would expect, that in a democratic society where government transparency and the fundamental importance of civil liberties dictates limits on the state's ability to take the lives of its citizens, that record-keeping of state-inflicted deaths would be meticulously collected and curated. [9] However, in both cases, the data that is publicly available through the federal government is significantly incomplete. [17, 19, 48, 54] For example, the *Washington Post*'s project to collect data on police killings for 2016 claims to have more than twice the number of such deaths than are recorded in FBI statistics. [53] Other projects with the same goal have found similar results. [26, 62]

Journalists have taken the lead in finding ways to collect national databases of police-inflicted deaths and to create easy ways for the public to access those resources. [17, 26, 48, 53] They are doing so using some of the tools of big data, such as combing the internet for daily references to such events, then plugging that data into ongoing projects. Last year alone there were over 1,000 killings in the U.S. by police. [26] With that data, other types of analysis can be done, such as mapping, and statistical models that relate predictor variables to differing rates of law-enforcement fatal violence. Since journalist efforts have produced, in most cases, the exact address where killings have occurred, the possibility of geocoding each event for advanced geospatial analysis is possible. [19] Similarly, since the geographic locations are known, other types of data, such

as demographic, economic and political data about the communities, or organizational information about the police department in question, can be used as independent variables. [36, 63]

While it is not known whether there has been an increase in citizen deaths caused by police in recent years, because of the unreliability of data about these incidents, there has been an increased awareness in the general population of the extent of these killings, arguably brought to the public's attention by the availability of social media, and easy access to filming and photographing equipment through the cameras on smart-phones, carried by a significant percent of the public. [7, 9, 48] Regardless, the issue of police brutality and particularly police killings has inspired movements such as Black Lives Matter to specifically highlight these issues, not only for people of color, but for all citizens. While these efforts largely became national movements after the killing of Michael Brown by police officer Darren Wilson, in Ferguson, MO, in 2014, scholars had already been looking into how to understand the extent of these killings, along with other types of excessive force by police. [14, 28, 44, 58, 63, 64]

The success of these movements in facilitating national dialogues about these issues, and available research on police brutality, create a foundation on which public policy might be constructed to minimize these outcomes, which are unique to the United States. Among all high-income countries, the US alone has rates of citizen deaths due to law enforcement above 0.25 people shot to death by police per million residents. While U.S. police killed 3.1 people per million, police in the next highest industrialized country, Australia, killed only 0.21 people per million, and Denmark was next, at 0.18 people per million. [60] According to reporting on this issue in *The Guardian*, police in England and Wales shot to death 55 residents in the last 24 years, while in the United States, police shot to death 59 residents in the first 24 days of 2015. [35]

## 2 POLICE USE OF EXCESSIVE FORCE, AND DEATHS CAUSED BY LAW ENFORCEMENT

The National Violent Death Reporting System, created by the Centers for Disease Control from a sample of seventeen participating states, tracks all reported violent deaths in those states. In 2013 they reported that about 1.2% of all violent deaths were caused by law enforcement, and non-Hispanic Black men were four times as likely as White men to be victims of law enforcement-caused deaths (as a percent of the population). [22] From multiple years of CDC data, this seems to be relatively consistent over the time of study.

Scholars, journalists and activists have noted both the rates of U.S. law-enforcement deaths compared to other industrialized countries, and the lack of reliable government data on the national number of

such killings. [17, 19, 35, 54, 60] While journalists, the FBI, and the CDC have been documenting basic counts and demographics of residents killed by police, scholars have tended to explore various theoretical frameworks that can make sense of these patterns. Most of these fall into a few categories: race-specific, individualistic, situational, organizational, and ecological (structural). [14, 48, 55, 64]

## 2.1 Individual, Organizational and Situational Theories

The individual approach argues there are specific characteristics of police that lead some to be more likely to use excessive force and kill suspects. Various characteristics have been studied, such as level of education, implicit bias (related to a race-approach, but on an individual level), professionalism, and personality traits. [48, 63] There is mixed to low support for any individual characteristic. Similarly, the organizational approach has mixed to low support. Organizational characteristics that are argued to be related to police use of excessive or deadly force are an increased and appropriately punitive bureaucracy for infractions, more stringent entrance requirements, higher levels of training, and greater levels of community accountability. [48, 63, 64] A counter-intuitive finding, is that higher levels of training can actually lead to higher levels of police violence, although there was little evidence to support a specific causal mechanism for this relationship. [64]

Situational theories propose that there is something specific in the interactions between police and citizens that led to the deadly force. Evidence suggests a positive relationship between victims who were non-compliant with police and/or impaired, and those killed by police. [48] A 2013 CDC report from 17 states where violent deaths were studied, describes that 39% of police-killing victims tested positive for alcohol (although no specific alcohol blood levels were reported), 18% tested positive for some form of opiate (just over half of victims were tested for opiates), 39% tested positive for marijuana (just over 40% were tested for marijuana), and 72% were armed. [22] However, that leaves a large number of victims who were neither impaired, nor armed. Not only is it troubling that non-impaired, unarmed citizens are being killed by law enforcement at such high rates, but it is also concerning that impaired or armed (not always with a gun) citizens face especially high risk for on-the-spot police-killing, when alternate approaches might have saved the victim's life. A case in point is Tamir Rice, the 12-year old who was shot by an officer within 2-seconds of arrival on the scene, because the boy had a realistic-looking toy gun. [48] Given that other industrialized countries are not facing these high rates of extra-judicial killings of suspects, it seems alternatives are viable. These issues continue to be raised by activist groups and community organizations.

While there is evidence for a higher risk of being killed by police when being non-compliant with law enforcement, citizen compliance is heavily dependent on how police treat them in a given situation. [11, 48] Police use of disrespectful and threatening tactics when approaching citizens is more likely to elicit disrespect and non-compliance from citizens, i.e., police aggression seems to cause non-compliance, not the other way around. [18, 41, 42, 61] There is mixed evidence whether people of color have higher rates

of non-compliance with police. Once class status is taken into account, one study found that Black men were *more likely* to be compliant than White men when confronted by police. [41] Regardless, evidence also suggests that the United States' history of police brutality against people of color, as well as over-policing specific to communities of color, significantly increases the likelihood that race minorities will be more suspicious of law enforcement, and thus potentially less compliant. [10, 34, 43, 51, 61]

## 2.2 Ecological/Structural Theories

Ecological/structural theories argue that levels of community violence, spatial organization, and community inequality are related to police violence. For example, evidence shows that in communities with higher crime rates, especially higher rates of homicide, police are more likely to use excessive force, and suspects are more likely to be killed by police. [48, 63, 64] Similarly, communities experiencing economic deprivation and inequality have been found to experience higher rates of police violence and police killings. [30, 36, 44, 51, 66]

Conflict theory is one of the primary structural approaches used to explain how state-constructed law enforcement systems seem to exploit and repress poor and minority communities. Rooted in the neo-marxist sociological tradition, conflict theory is a political-economic approach that acknowledges social power imbalances, and argues that conflict between those who hold power in society—those who hold political office, land-owners, the wealthy, and the race/sex/sexual identity majorities—deploy political and legal strategies to reduce threats, real or perceived, by other groups. [28–30, 33, 34, 52, 64] This strategy typically criminalizes those groups that are subject to these tactics, and thus acts as a moral-legitimizing cultural discourse to justify these structures to the public, thus reinforcing the repressive policies. Given that race minority communities have significantly lower income, wealth, and political representation, sociologists argue that it can be difficult to disentangle the impacts of race versus class, and thus intersectional approaches can be productive. [25, 38, 51]

Race-specific theories build on a long history of critical race studies documenting racial inequality in the United States, from slavery until today. [14, 16, 44, 51, 55] These approaches explore concepts like structural racism, implicit bias, discrimination, segregation, the prison-industrial complex, and economic disenfranchisement. In terms of interactions with the legal system, critical race theory argues that race minorities are often targeted specifically because of race, tracing back to the slave-era, to the era of lynchings, and to today where data indicates significantly higher rates of conviction and imprisonment for race minorities, and higher rates of the use of excessive force by police. [14, 16, 44, 51, 52, 55]

A specific variant of this approach explores the social psychological effect of inter-group threat perceptions. The minority threat hypothesis proposes that very small groups of race minorities pose little economic or political threat to the majority. However, once a critical threshold is reached, the majority perceives significant threat from the minority, and deploy various tactics to try to neutralize that threat, such as discriminatory housing policy, increased use of police as a tool for economic extraction and political violence, education discrimination, and employment discrimination.

[37] Conversely, a parabolic curve is predicted, such that once the minority population grows above 50%, they should have sufficient political power over time to counter these patterns. While some early research found only mixed evidence to support the minority threat hypothesis, the majority since have shown significant support. [28, 30, 34, 36, 52, 63, 64, 66] Further, from subsequent reanalysis of studies that did not show robust minority threat effects, it is argued that the earlier efforts examined raw population numbers, but once proportionate shares of the population were used instead, significant effects were found. [36]

### 3 TRACKING DEATHS CAUSED BY LAW ENFORCEMENT

There is no reliable central database in the United States for any killings, whether caused by citizens or law enforcement. The best publicly available government resource is the FBI database on homicides from the Uniform Crime Report (UCR). However, as numerous sources have documented, this database is unreliable because participation by local and state law enforcement agencies is voluntary. [17, 19, 54] There have been several attempts by researchers and law enforcement to get a better estimate of homicide rates. For example, the FBI's Supplementary Homicide Report, and National Vital Statistics Systems are two federal-level attempts to collect detailed statistics, but those efforts are not presumed to represent the entire population of homicides. [54] Similarly, the Centers for Disease Control collected a sample database, originally of just sixteen participating states in 2005, the National Violent Death Reporting System, cataloguing various aspects of all reported violent deaths, including whether law enforcement was involved. [22] Each of these is not a full-scale national survey, but non-random samples from the population. However, what they all show is the failure of the UCR to capture even a majority of law enforcement-caused deaths.

Data on homicides specifically reported to have been committed by law enforcement face similar deficits. [17, 48, 54] To overcome this problem, journalists, academics, and activist organizations have used publicly available sources, primarily news reports, to find instances of law enforcement-caused deaths. For example, the British news agency, *The Guardian*, created a two-year project, *The Counted*, where they manually searched news reports for cases of deaths caused by law enforcement. [26] Additionally, they posted contact information for the public to send them tips about cases not already in their database. While *The Guardian* database for 2015–2016 remains available to the public, they discontinued the project, so no 2017 data is available. It has some interactive features, where the user can filter the data by state where the killing occurred, how the death occurred, whether the victim was alleged to have been armed, victim's gender, race, and age. Where possible, a picture of the victim and a brief biography is available. They report that 1,093 people were killed by law enforcement in 2016 alone.

Another news agency that relies on a similar process to capture data about deaths caused by police, is the *Washington Post*. They host an ongoing project that lists only people shot to death by police, Fatal Force, and contains links to the original news stories where the deaths were reported. [53] Its numbers are lower than *The Counted*, which includes any deaths caused by law enforcement,

not just shootings. However, it also notes that its database still has more than twice the number of shootings by law enforcement per year than the FBI database. Like *The Guardian*, it has contact information for the public to send new information about killings, as well as photos or videos about the victims. Their list of victims starts from January 1, 2015. They report that 963 people were shot to death by law enforcement in 2016.

*Mapping Police Violence* (MPV) was created by three community activists and organizers. Their reported methodology is to have used three online databases, and to have compiled those lists into MPV. Neither *The Guardian* nor the *Washington Post* are reported as a source for MPV. They operationalize police killings as, a 'case where a person dies as a result of being chased, beaten, arrested, restrained, shot, pepper sprayed, tasered, or otherwise harmed by police officers, whether on-duty or off-duty, intentional or accidental.' [62] Their list of victims begins from January 1, 2013. They report that 1,155 people were killed by law enforcement in 2016. This is the database used for this project to identify the counties where people were killed by law enforcement.

### 4 CRITICAL STUDIES, COUNTER-MAPPING, AND COUNTER-DATA

In addition to efforts at tracking law enforcement-related deaths, a broader movement has arisen in the era of big data, merging the scholarship and practices of critical studies with data science. Specifically, some scholars and activists are discussing practices of counter-data and counter-mapping, as an attempt to mitigate the state's use of big data and mapping to enlarge the carceral state. [15, 17, 19, 31] This approach highlights how government and capitalism construct a specific vision of the world that creates tight boundaries around civil liberties, rather than expanding civil liberties, which is arguably one of the original visions of a good society under liberalism.

#### 4.1 Critical Studies, Foucault

Critical studies began in the 1920s at the Institute for Social Research in Frankfurt, Germany, in response to the rise of fascism. [57, 59] Several key thinkers, such as Adorno, Horkheimer and Marcuse, tried to integrate neo-Marxism, Max Weber, and Freud, into a liberating sociology that rejected the growth of the oppressive state. All of these early founders wrote and spoke vigorously against Hitler, and were eventually forced to flee Germany, transplanting the critical studies movement to Columbia University. [57] Their outlook critiqued the current (at the time) emphasis in the social sciences of positivism, which seemed to strip *humanity* out of social ideas, reducing people to the condition of biological robots. Further, the growing efficiency and expansion of the state with the success of bureaucracy and the so-called 'scientific management' of Frederick Taylor, society was increasingly treating people like the robots described by sociology and psychology. Finally, the successful takeover by capitalism of all social processes, not only the economy, but a restructuring of how humans relate to each other—as competitors, as sources of capital, and bodies to be exploited for labor—brought critical studies to critique the entire direction of society. This approach contrasted with the Weberian, and *scientific*

approaches that argued for neutral objectivity as a social scientist. [57, 59]

Subsequent scholarship built on these ideas. Michel Foucault, a French scholar writing in the 1960s and 70s, produced a series of critiques of social structures, from an examination of the growth of prisons, to problematic psychiatric approaches at mental hospitals, to how cultural approaches to sexuality were often used as a specific form of social control and repression. In contrast to the original critical theorists who built on Marx and Freud, Foucault rejected both of those approaches to understanding society. Two of Foucault's methods he describes as archaeology of knowledge, and genealogy of power. [57, 59] Much as traditional archaeology is a process of digging into the past to see how an ancient society was built, Foucault's archaeology of knowledge was a process of digging into texts to discover how knowledge systems are built that then structure how various social discourses are deployed, that then dictate all social arrangements. Similarly, much as a family genealogy traces back how individual people are related to each other, Foucault's genealogy traces back ideas to see how they are related to each other, and in doing so, describes ways to understand power in society, since knowledge and information in industrial and post-industrial societies are a key to power consolidation.

Foucault argues that if the leaders of a society can harness technologies to maximize their knowledge of members of that society (which they do), they can wield great power over those members. For example, in *Discipline and Punish*, Foucault describes the concept of Panopticon, based on a prison design by Jeremy Bentham in 1791. Built in such a way so that the prisoners never knew when they were being watched by the guards, and they could easily be seen at all times, the idea is that they would eventually start to subjugate their own behaviors to the will of the prison guards, thus minimizing the efforts the prison had to take to control them—the prisoners disciplined themselves rather than having to be constantly disciplined and punished by the guards. [23] Similarly, modern school rooms are typically structured such that the teacher stands at the front of the room, while students sit at desks that are arranged in rows facing forward. The teacher can easily observe all students in one glance, and the students know they can be seen at any given moment; thus are more inclined to discipline their own bodies. Foucault argues that by simply creating a technology of ubiquitous observation, it radically reshapes citizen behaviors, limiting freedom of expression. These issues became part of daily life under Stalin and Hitler, as hordes of citizens were conscripted into the government machinery to be spies on their neighbors. All aspects of citizens' daily lives were monitored, reported, documented, and if applicable, punished.

## 4.2 Foucauldian Activists, Counter-Data

Contemporary activists and scholars recognize the ways that Foucault's vision of a ubiquitous monitoring system can constrain human liberty. Jefferson highlights this concern in a description of the Chicago Police Department's use of digital mapping for crime, CLEARmap, in a cleverly-titled article based on Foucault's essay, *Digitize and Punish*. [31] Jefferson describes the fact that CLEARmap is not a neutral, objective mapping system, but that it

'provides ostensibly scientific ways of reading and policing negatively racialized fractions of surplus labor in ways that reproduces, and in some instances, extends the tentacles of carceral power.' In doing so, Jefferson argues for the linking of critical mapping ('critGIS'), and critical race/ethnic studies. Because of how highly over-policed race-minority communities are, these are the regions most likely to have high concentrations of citizens pulled into the carceral state. [10, 11, 14, 16, 25, 29, 43, 51, 56, 58, 61] Thus, such mapping not only allows for the visualization of crime, but it also becomes a tool for racial subjugation, expanding police and state gaze into, and thus actions upon, communities already over-policed. Brayne explores the same concerns and effects in a study of the expansion of big data surveillance by the Los Angeles Police Department. [8]

Currie, *et al*, summarize the field of counter-data, which incorporates critical studies with big data. [17] Specifically, concerned about the lack of transparency of the various, overlapping layers of government as it relates to police brutality in communities, they argue that residents should use the tools available to them to construct their own data sets, thus creating leverage for greater government accountability. This requires creating infrastructures for gathering, processing and distributing information to the public. They note that government data on police brutality is widely recognized as flawed and grossly incomplete. [17, 19, 50] Noting that government data collection can be easily used to constrict citizen liberties, while at the same time, failing to protect citizens from state violence, they point to counter-data as a process by which citizens collect their own data as a means of protection and activism.

Counter-mapping is an extension of counter-data, but applies the process of data-collection to how people think about the spatiality of the places around them. Dalton and Stallman describe counter-mapping, which, 'involves map-making practices by those outside or on the margins of large, powerful institutions such as corporations or governments. The modern history of most maps and GIS is one of government programs, such as extending territory, military conflicts, property cadastres, or administering territories' (p. 3). [19] They point to activist mapping projects, such as *Mapping Police Violence*, as a way for citizens to map how the state imposes power unjustly, and package it into formats that are readily available to the public, thus empowering them as citizens by increasing their awareness of state practices. Similar citizen-activist mapping projects to oppose state imposition of power are described by Maharawal, regarding gentrification in San Francisco, and the Counter Cartographies Collective, *et al*, describing counter-mapping as 'militant research' that benefits the masses overtaken by an overreaching, exploitative state. [15, 38]

## 5 DATA AND MEASURES

### 5.1 Dependent Variable: Count of People Killed by Police

The dependent variable for this study is a count of the number of people killed by police in any given county in the United States from January 2013 to October 2017 and was downloaded from *Mapping Police Violence* on November 5, 2017. [62] Since the MPV site is constantly being updated, the original data downloaded is available from the author's Google Drive course site. [69] Compared to

the two news agencies that have been tracking deaths from law enforcement, their list of victims is more comprehensive for 2016 (1155 vs 963 and 1093), and it covers a longer time-frame, since the news agencies only have data since 2015. The original xlsx (Excel) file contains 24 variables on the primary sheet, such as the victim's name, whether the victim was armed, a link to the original news story, whether there were reports of mental illness, the county, zip-code, state, specific address of the killing, victim's age, race, gender, date of the incident, and police department involved in the killing. At the time of download, the spreadsheet contained a total of 5,634 observations, that each represent a killing by law enforcement in the U.S. in approximately the last five years.

This data contained errors that had to be manually corrected. The errors were discovered when the "killed" variable, count of the number of people killed by law enforcement in any given county, was merged with Census data, the independent variables. The Census data contains an official list of counties in every state. The MPV data listed counties that were not in the claimed states. It was discovered that some were simply errors, some were cities coded as counties, and some counties were blank. Sixty observations total were found to have errors. Using the victim's name provided by MPV, they were searched using Google (November, 2017), to find an original news story reporting this killing. The pre-cleaned data is available on the author's Google Drive course site. [69] Once the killing could be verified, similar investigation confirmed the county and state of the killing. The corrected data was used for this analysis, and is available from the author's Github course site. [71] Every missing or incorrect county was found. It is possible that other errors exist, since the only errors detected were those where there was no correct match with an actual Census-listed county and state. There may be counties where a victim is claimed to have been killed, where the county and state exist, but misidentified. The data has approximately the same number of observations as the two news agencies' data, strengthening its claims to validity. Further, each of the 60 misspecified counties that were corrected, traced back to an actual killing by law enforcement based on a search of the victim's name, increasing confidence in the data.

## 5.2 Independent Variables: Census Data

Summary statistics and definitions can be found in Table 1 for all of the variables used. The independent variables were all obtained from the United States Census, which maintains a large, publicly-available database on their website. [12] All of the data used for this study are at the administrative unit of the county, and includes all of the 3,142 United States counties listed by the Census. All of the ten variables were downloaded (November, 2017) from the American Community Survey (ACS). The ACS is a subsidiary of the Census Bureau which administers an extensive survey of an annual sampling of the U.S. population on various economic and demographic questions. Each of the variables chosen for this project falls into one of five categories: population total, race, education, geography-type and economic. The total population of each county is estimated by the ACS based on decennial census data, and represents a single-year value. The rest of the variables are estimated from aggregates of data collected from 2011-2015, a five-year estimate. Multiple studies have shown that for homicide analysis,

whether resident-on-resident killings, killings by police, or killings of police, demographic and economic variables tend to have strong predictive value. [33, 36, 50, 54, 64]

[Table 1 about here.]

The only race variable in this study is a ratio of White:Black population per county. The Census uses self-identification of the individuals to label race. Individuals can self-identify as multiple race and ethnic groups. For this study, only those individuals listing a single race, White or Black, were counted, and a ratio was calculated. The geography-type variable represents whether a person lives in a rural or urban area. The Census defines *urban* as any place with at least 2,500 people, and *rural* as everyplace else. The geography-type variable used for this study is a percent of the county population who does not live in an urban area. The two education variables used, only count those residents who are over 25 years of age, either who have a bachelor's degree or higher, or who do not have a high-school diploma.

Five economic-related variables were used. The first, percent of people employed per county, is taken only from those 20-64 years of age. The second, measures the median income per county. Income can be measured in several different ways, and while mean is a common measure of central tendency, it is often skewed up by high incomes, since low incomes are limited to zero. Because of this, median income is more often used. Two variables measure income deficits—the percent of people who qualified for SNAP (food stamps), and the percent of people living below poverty. The latter relies on the definition of poverty provided by the Office of Management and Budget, which varies this designation based on size and composition of the family. The fifth measure is the percent of the population living above \$100,000 per year. The Census lists the median U.S. income at almost \$60,000 per year, and provides the \$100k data as a measure of a high-income county.

## 5.3 TIGER Shapefiles for Mapping

In addition to demographic and economic data, the Census Bureau provides a significant amount of geospatial information. Through their TIGER products (Topologically Integrated Geographic Encoding and Referencing), they provide a web site where the public can download shapefiles at many different levels that are published each year, and can be connected to the rest of their data. [13] For this study, county-level data was used, since it was the smallest administrative unit for which all of the data was available. While accurate state-level data was available, it did not seem fine-grained enough to provide a sufficient analysis. Smaller administrative levels, like census-tract, and even block-level shapefiles are available, and there is annual census data that can be mapped onto those levels. While some ACS variables are available at these small units, they typically only cover a random selection, and can have high error rates. County-level data is the smallest administrative unit for which every county in the U.S. can be measured in the five-year estimates file. The shapefile used for this study is the 2015 county-level data for the entire United States, and was downloaded from the Census TIGER site (November, 2017). [13] However, not all places available on the shapefile have data through the ACS. For example, several of the smaller island territories are not included in the ACS survey, so those are excluded in the map, as well as the

regression analysis. While the latter includes Alaska and Hawaii, for ease of viewing the county-level data, those two states are not included on the map (Figure 1).

## 6 METHODS

There are multiple ways to attempt to understand police killings of citizens. One way is by visualizing data of the events. Another is by creating statistical models that relate the killings to predictor variables, such as in a regression equation. Open source software is available to facilitate both of these approaches.

### 6.1 Mapping, Quantum GIS

Quantum GIS (QGIS) is open source software that processes the visualization of spatial information, and is useful for tasks such as mapping, and seeing how data are related in geographic ways. [3] For this project, QGIS 2.18.8 was used to process the Census TIGER shapefile (shp) that contained all of the United States counties. When downloading files from the TIGER site, several other files come with the shp file, one of which is a database file (dbf) that contains information about each place identified in the shapefile. For example, land and water area frequently come with the dbf, along with government geocodes and names that identify each specific location, in this case, every county in the United States.

Since the data used here is coded into a county-level datafile, the first entry in the dbf is Cuming County, Nebraska, with Nebraska geocoded as state #31, and Cuming County geocoded as county #039. All 3,142 counties in the dbf file have these coded identifiers that are matched to demographic and employment data that can be downloaded separately from the Census Bureau, such as ACS data. Since the geocodes are identical across all data sets for each specific geography, this data and the shapefiles can be readily integrated. Integrating other types of data, such as the police killings data from MPV, requires a different approach. In this case, since MPV provided both the county and state names where the deaths occurred, merging these into one string, such as "CumingNE", and doing the same with the TIGER shapefile data, allows the two files to be jointly identified and merged. For Figure 1, which shows the rates of killings by police at the county-level, a ratio was calculated in QGIS—the count of those killed in each county divided by the total population of that county for the measure, rates people killed by police per 100,000. With the creation of this new variable, it can be mapped in QGIS as a *graduated symbol* with a pre-specified color scheme.

[Figure 1 about here.]

This process can be automated with Python, a programming language useful for big data analysis. [2] QGIS has integrated Python into its software with the inclusion of a Python Console for direct use, and an online instruction manual available through the QGIS web site. [1] For the map in Figure 1, a Python script was created that downloads a zipped file containing the TIGER shapefile for the US counties, and all of the census data that had previously been integrated into the dbf file. The script and data are available on the author's github course site. [70] The zipped shapefile is available on the author's Indiana University Box account (it was too large for Github to allow). [68] The Python program unzips the file, deletes all parts of the map except for the contiguous United

States (all but 48 states), then applies a graduated color scheme to the killed per population variable. Figure 5 shows part of the code to download and extract the data.

### 6.2 Regression, Count Data and Data with a Mass at Zero

Regression analysis can be performed, and graphs created, through open source software, R, an environment for statistical analysis and graphics. [4] In addition to the base R package, which is command line only, other packages are available as an overlay to incorporate GUI features, such as RStudio. [5] This analysis was done with R 3.4.1, using RStudio 1.0.153. Instead of Python, R has its own scripting language. The R-script for the following procedures and data are available on the author's github course site. [71] Figure 6 shows the code to download and extract the data.

Creating a statistical model is a way to understand how variables are related to each other. Regression analysis is a common way to model such relationships. While often associated with a continuous dependent variable, regression can also be used to model count variables. [33] In this case, the number of people killed by police is a count variable, and all of the predictor variables are continuous. Modelling continuous dependent variables presumes a Gaussian distribution, but count variables often do not meet this assumption. [24] While transforming a non-normal dependent variable is the typical procedure to get it into a state of normal distribution for analysis, using square root, reciprocal, log, or even a Box-Cox procedure, count variables can resist this process. Poisson and negative binomial distributions are typically better suited for count data. [40] These distributions can have an unusually large mass at zero, one of the prime reasons transformations often fail to produce a normal distribution. [6, 20, 24, 47, 65]

Count data with a mass at zeros can be analyzed in a number of ways, depending on the theoretical reason for the zeros, each of which requires a different analytic approach. One decision factor includes knowing whether there is one process generating all of your data, or whether there are two—one generating the count data (above zero) and some of the zero, and a second process generating *an excess* of zeros. [20, 40, 45, 47] In the latter case, these zeros would be more than would be predicted from the given distribution, and those are sometimes called *inaccessible* zeros. Inaccessibility zeros are considered excessive, and a two-part, zero-inflated model might be a good option in that case. Zero-inflated Poisson, and zero-inflated negative binomial models are available in R, in packages such as *gamlss* and *pscl*. [65] In Figure 2, the *killings by police per county* dependent variable is shown in a histogram, and as a plot against the county population, both graphs being created in R using the standard *hist* and *plot* functions. This allows one to visualize the large mass at zero, which might lead one to presume a zero-inflated model is most appropriate.

[Figure 2 about here.]

However, in the case of police killings data, a zero-inflated model is not theoretically sound. Such models presume that there are a group of 'zero count' observations that are zero because it is impossible for that member to be anything other than zero. For example, if there were no police in a given county, then there could be no police killings there, and thus, would represent a different

*zero-generating process* than that creating the rest of the data. [21, 45, 47] A health care example, is that if you were counting people in a given county who accessed health care services, and there was a group of people who lacked access to health care, they would be counted as zeros not because they did not need health care, but because they could not access it. In this case, there are two separate processes generating the data. [47] Tetzlaff describes analyzing county-level child homicides: since every child in a county has a possible chance of being killed (fortunately most are not), a zero-inflated model would not be appropriate. Since there are no children for whom it is impossible to not become a homicide count, a zero-inflated model is theoretically unsound. [67]

Other possible models are available for data with a mass at zero, for example, a two-part hurdle model. However this approach would also be inappropriate because of hurdle assumptions. For example, it presumes that all cases for whom becoming a count is possible, it will become a count—in other words, if criteria are met, then the subject would always move over the hurdle from a zero to a count. [21, 46] Tobit models are often used with continuous data and a mass at zero. But a Tobit zero is assumed to be possible censored data, such that it is some value other than an actual zero. In this case, where a count of police killings is the dependent variable, a zero represents a true zero, not a possible censored zero, so a Tobit model is also likely not a good choice. [45, 47]

### 6.3 Poisson and Negative Binomial Distributions in R

One of the assumptions of a Poisson distribution is that the mean and dispersion are equivalent. A model with a variance greater than the mean is considered over-dispersed, which implies it is a better candidate for a negative binomial regression approach. [6, 24, 36, 64] As can be seen in Table 1, the mean of the count killed by police is 1.74, and the variance is 57.0 (square of standard deviation, 7.55). This data is clearly over-dispersed, and is thus a good candidate for negative binomial regression.

Since neither Poisson nor negative binomial meet the data assumption of a normal distribution, alternate methods of analysis are required beyond standard linear regression. [6, 20, 24] In these cases, general linear models (GLM) can be used for the analysis, and is considered a form of nonlinear regression that uses one of the exponential distributions. [24, 32, 45] Whereas linear regression is composed of a random element and a linear or systematic element, GLMs have a third element, a smooth and invertible link function, such as log, inverse, square-root, or probit, which transforms the response variable. [24] GLM has more flexibility in how means are estimated, such as maximum-likelihood (ML) estimation rather than ordinary least squares (OLS). The parameters for the *killings by police per county* data here is estimated using GLM, maximum-likelihood, a log-link function, and a negative binomial distribution assumption.

Several packages in R are available for this type of analysis, such as MASS::glm.nb, mcgv::gam, and gamlss::gamlss. The latter is used here for the primary analysis, because of the ease of plotting residual diagnostics, and the simplicity of constructing the command, compared to the other two packages. However, all three approaches are included in the R-script (available at the author's github site)

to compare results, and to produce specific goodness-of-fit results. [71] The package, gamlss, automatically estimates all relevant parameters given just the dependent and independent variables, but only provides AIC as a way to compare models. [65] Similarly, glm.nb provides AIC, but also produces the log-likelihood. The latter provides an estimation of the dispersion (shape) parameter  $\theta$ , required for a negative binomial model, which gamlss will not provide. For an additional goodness-of-fit estimation, gam is used to estimate the parameters of the negative binomial model, since it provides an estimation of both the R-squared, and deviance explained. [39] However, gam does not automatically estimate  $\theta$ , so that value is estimated from glm.nb, then passed to gam for this analysis. The estimates for the coefficients for all three approaches are virtually identical, although there were slight, but non-significant, variations in the p-value estimates.

## 7 FINDINGS

Three regression models were created: the full model, which includes all ten independent variables; the partial model, which includes only five independent variables; and the final model, which includes only the three predictors with the strongest p-values. The coefficients and p-values of these models can be found in Table 2, along with a number of goodness-of-fit and diagnostic information. From the full model, all of the non-significant variables were removed to generate the partial model. From the partial model, the two variables whose p-values were greater than 0.05 were removed from the model, leaving just three predictor variables, all of which were significant to  $p \leq 0.001$ . [6] By all measures, the three models have very similar model fit and diagnostic criteria (AIC, R-square, correlation between predicted and observed, RMSE, predicted mean of outcome variable, etc.). The variance inflation factor was checked for each, and while significant problems were detected for the full model ( $\max VIF = 34.5$ ), those problems were eliminated in both of the smaller models. The final model is arguably the best model since it is the simplest, with only three predictor variables, and having equivalent goodness-of-fit and diagnostic values as the more complex models.

[Table 2 about here.]

The final regression model indicates a strong relationship between the dependent variable, the count of people killed by police in each county, and the three predictor variables, population of the county, employment rates in that county, and county median income. The coefficients indicate both strength and direction of the relationship, and can be used to create a prediction equation.

$$\text{Killed by Police} = -10.01 + 1.092 * \log(\text{Population}) - 0.0185 * \text{Employment Rate} - 0.000015 * \text{Median Income}$$

The coefficients imply that the higher a county's population, the higher the risk of being killed by police; the higher the employment rate in the county, the lower the risk of being killed by police; and the higher the county median income, the lower the risk of being killed by police. The coefficients are interpreted as the change in the log-count killed by police for each unit change of the independent variables, presuming the other variables remain the same. For practical purposes, this implies that taking the exponent of the

predictor coefficient allows one to interpret the coefficients like a normal linear regression. For example, as employment in a county rises by  $\exp(0.0185)$ , or 1.02% it predicts a decrease of one count of police killing in that county.

The variance inflation factor did not indicate multicollinearity between these variables (max VIF=1.99). [33, 48] According to output from the R regression analysis, this three-variable model explains 76.6% of the variance of the dependent variable, and the predicted vs observed killings by police have a correlation of 0.936. The mean predicted killings are 1.78 per county, while the observed killings were 1.74. This model predicts that 1,656 counties will have approximately zero killings (less than 0.5, not exactly zero), while the actual number of counties with no killings was 1,827. [6]

Figures 3 shows two graphs supporting the conclusion that the final model is an acceptable fit. The left plot of predicted versus observed killings by police per county, with the red regression line of just these two variables, shows the strong relationship between these two variables. The rootogram on the right shows that few bins of predicted outcomes of the final model are significantly different from the observed values. A rootogram is interpreted by the *hanging bars*—bars hanging above the zero-line are over-fit, while bars hanging below the line are under-fit. [49] The closeness of the bars to the zero-line indicate a relatively good fit.

[Figure 3 about here.]

In addition to goodness of fit, regression also requires that residuals meet basic assumptions. Figure 4 shows several standard diagnostic plots indicating that these assumptions seem to be met. The residuals versus fitted values plot is interpreted as being problematic if a pattern emerges from the residuals, either above or below the zero-line. In this case, no pattern is apparent—the residuals seem equally and randomly scattered above and below the line. Similarly, the plot of residuals versus index shows a similar result, indicating the regression residuals assumption is met. Finally, the bottom two plots, more residuals plots, show the same. For example, the Normal Q-Q Plot would indicate a problem if a significant number of dots were straying from the main central line—in this case they are almost all on the line. [27]

[Figure 4 about here.]

## 8 CONCLUSION

The United States has an inordinate number of killings by law enforcement compared to other high-income countries. People of color face the brunt of this violence as a share of the population. Several theories have been proposed to explain these problems, such as minority threat, community inequality, police training, and citizen non-compliance. One of the difficulties researching police violence is the lack of reliable data available to the public. Until recently, the primary source for data was the FBI's Uniform Crime Report, which is widely known to be severely flawed. Efforts by journalists and community activists have created several resources that are available to the public, and can be used for research on these types of killings.

Demographic, economic, political, and organizational variables can be analyzed with regression methods to model the types of social factors that are related to state violence. A number of researchers have consistently found certain factors, like poverty, race

and population size, to be related to levels of police violence. Because counts are not normally distributed variables, negative binomial approaches have frequently been used for this type of data, and have produced consistent results. Since the data that is available has been coded to many factors, including the specific location of the killing, mapping these deaths are possible.

This project used mapping in QGIS to show county-level rates of violence in the United States, and regression in R to model the relationship of several variables to killings by police, finding that the strongest, yet simplest model predicting police killings are the population size, employment rates, and median income, with the latter two being negatively related to killings by police. This study did not look at any variables that could be considered individual, organizational, or situational, three theories commonly used to explain police violence. However, the fact that two economic variables were significant predictors of deaths at the hands of law enforcement, would seem to contribute to the ecological approach, specifically, theories of community economic structural factors.

## REFERENCES

- [1] 2017. PyQGIS Developer Cookbook. online. (2017). [https://docs.qgis.org/2.14/en/docs/pyqgis\\_developer\\_cookbook/](https://docs.qgis.org/2.14/en/docs/pyqgis_developer_cookbook/)
- [2] 2017. Python. online. (2017). <https://www.python.org/>
- [3] 2017. Quantum GIS. online. (2017). <http://www.qgis.org/en/site/>
- [4] 2017. The R Project for Statistical Computing. online. (2017). <https://www.r-project.org/>
- [5] 2017. RStudio. online. (2017). <https://www.rstudio.com/>
- [6] Alexander Beaujean and Grant B. Morgan. 2016. Tutorial on Using Regression Models with Count Outcomes using R. *Practical Assessment, Research and Evaluation* 21, 2 (2016), 1–19.
- [7] Yariimar Bonilla and Jonathan Rosa. 2015. Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist* 42, 1 (2015), 4–17.
- [8] Sarah Brayne. 2017. Big Data Surveillance: The Case of Policing. *American Sociological Review* 82, 5 (2017), 977–1008. <https://doi.org/10.1177/0003122417725865>
- [9] Ben Brucato. 2015. The New Transparency: Police Violence in the Context of Ubiquitous Surveillance. *Media and Communication* 3, 3 (2015), 39–55.
- [10] Rod Brunson and Jody Miller. 2006. Gender, Race, and Urban Policing: The Experience of African American Youths. *Gender and Society* 20, 4 (2006), 531–552.
- [11] Rod K. Brunson and Jody Miller. 2005. Young Black Men and Urban Policing in the United States. *The British Journal of Criminology* 46, 4 (2005), 613–640.
- [12] United States Census. 2010–2015. American Community Survey. online. (2010–2015). <https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml>
- [13] United States Census. 2015. Tiger Shapefiles. online. (2015). <https://www.census.gov/geo/maps-data/data/tiger-line.html>
- [14] Cassandra Chaney and Ray V. Robertson. 2013. Racism and Police Brutality in America. *Journal of African American Studies* 17, 4 (2013), 480–505.
- [15] Counter Cartographies Collective, Craig Dalton, and Liz Mason-Deese. 2012. Counter (Mapping) Actions: Mapping as Militant Research. *ACME: An International E-Journal for Critical Geographies* 11, 3 (2012), 439–466.
- [16] Wesley Crichlow. 2014. Weaponization and Prisonization of Torontofis Black Male Youth. *International Journal for Crime, Justice and Social Democracy* 3, 3 (2014), 113–131.
- [17] Morgan Currie, Britt S Paris, Irene Pasquetto, and Jennifer Pierre. 2016. The conundrum of police officer-involved homicides: Counter-data in Los Angeles County. *Big Data and Society* (2016), 1–14.
- [18] Mengyan Dai, James Frank, and Ivan Sun. 2011. Procedural justice during police-citizen encounters: The effects of process-based policing on citizen compliance and demeanor. *Journal of Criminal Justice* 39 (2011), 159–168.
- [19] Craig Dalton and Tim Stallman. 2017. Counter-mapping data science. *The Canadian Geographer* early view, online only (2017). <https://doi.org/10.1111/cag.12398>
- [20] VT Farewell, DL Long, BDM Tom, S Yiu, and L Su. 2017. Two-Part and Related Regression Models for Longitudinal Data. *Annual Review of Statistics and Its Applications* 4 (2017), 283–315.
- [21] Cindy Feng and Longhai Li. 2016. *Advanced Statistical Methods in Data Science*. Springer, Chapter Modeling Zero Inflation and Overdispersion in the Length of Hospital Stay for Patients with Ischaemic Heart Disease, 335–53.

- [22] Centers for Disease Control. 2016. Surveillance for Violent Deaths – National Violent Death Reporting System, 17 States, 2013. online. (2016). <https://www.cdc.gov/mmwr/volumes/65/ss/ss6510a1.htm>
- [23] Michel Foucault. 1995 (1977). *Discipline and Punish: The Birth of the Prison*. Vintage Books.
- [24] John Fox. 2015. *Applied Regression Analysis and Generalized Linear Models*. Sage.
- [25] Keon L. Gilbert and Rashawn Ray. 2015. Why Police Kill Black Males with Impunity. *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 93, Supplement 1 (2015), S122–S140.
- [26] The Guardian. 2015-2016. The Counted. online. (2015-2016). <https://www.theguardian.com/us-news/series/counted-us-police-killings>
- [27] Daniel Hocking. 2011. Model Validation: Interpreting Residual Plots. online. (2011). <https://www.r-bloggers.com/model-validation-interpreting-residual-plots/> Retrieved Nov 21, 2017.
- [28] Malcolm Holmes. 2000. Minority Threat and Police Brutality: Determinants of Civil Rights Criminal Complaints in U.S. Municipalities. *Criminology* 38, 2 (2000), 343–68.
- [29] Malcolm D. Holmes and Brad W. Smith. 2012. Intergroup dynamics of extra-legal police aggression: An integrated theory of race and place. *Aggression and Violent Behavior* 17, 4 (2012), 344–353.
- [30] David Jacobs and Jason Carmichael. 2002. Subordination and Violence against State Control Agents: Testing Political Explanations for Lethal Assaults against the Police. *Social Forces* 4, 1 (2002), 1223–1251.
- [31] Brian Jefferson. 2017. Digitize and punish: Computerized crime mapping and racialized carceral power in Chicago. *Environment and Planning D: Society and Space* 35, 5 (2017), 775–796.
- [32] Andrew Jones. 2011. Models for Health Care. In *The Oxford Handbook of Economic Forecasting*, Michael P. Clements and David F. Hendry (Eds.). Oxford University Press, 625–654.
- [33] Robert Kaminski. 2008. Assessing the County-Level Structural Covariates of Police Homicides. *Homicide Studies* 12, 4 (2008), 350–380.
- [34] Stephanie Kent and David Jacobs. 2005. Minority Threat and Police Strength from 1980 to 2000: A Fixed-Effects Analysis of Nonlinear and Interactive Effects in Large U.S. Cities. *Criminology* 43, 3 (2005), 731–760.
- [35] James Lartey. 2015. By the numbers: US police kill more in days than other countries do in years. online. (June 2015). <https://www.theguardian.com/us-news/2015/jun/09/the-counted-police-killings-us-vs-other-countries>
- [36] Joscha Legewie and Jeffrey Fagan. 2016. Group Threat, Police Officer Diversity and the Deadly Use of Police Force. *Columbia Law School Public Law and Legal Theory Working Paper Group* 14-512 (2016), 1–42.
- [37] Joseph Luders. 2010. *The Civil Rights Movement and the Logic of Social Change*. Cambridge University Press.
- [38] Manissa M Maharawal. 2017. Black Lives Matter, gentrification and the security state in the San Francisco Bay Area. *Anthropological Theory* 17, 3 (2017), 338–364.
- [39] Jacob Martin and Daniel B. Hall. 2016. R2 measures for zero-inflated regression models for count data with excess zeros. *Journal of Statistical and Computation and Simulation* 86, 18 (2016), 3777–3790.
- [40] Jacob Martin and Daniel B. Hall. 2017. Marginal zero-inflated regression models for count data. *Journal of Applied Statistics* 44, 10 (2017), 1807–1826.
- [41] Stephen Mastrofski, Jeffrey Snipes, and Anne Spina. 1996. Compliance on Demand: The Public's Response to Specific Police Requests. *Journal of Research in Crime and Delinquency* 33, 3 (1996), 269–305.
- [42] John McCluskey, Stephen Mastrofski, and Roger Parks. 1999. To Acquiesce or Rebel: Predicting Citizen Compliance with Police Requests. *Police Quarterly* 2, 4 (1999), 389–416.
- [43] Albert Meehan and Michael Poncer. 2002. Race and Place: The Ecology of Racial Profiling African American Motorists. *Justice Quarterly* 19 (2002), 399–430.
- [44] Daryl Meeks. 2006. Police Militarization in Urban Areas: The Obscure War Against the Underclass. *The Black Scholar* 35, 4 (2006), 33–41.
- [45] Yongyi Min and Alan Agresti. 2002. Modeling Nonnegative Data with Clumping at Zero: A Survey. *Journal of The Iranian Statistical Society* 1, 1 (2002), 7–33.
- [46] Will Moore and Stephan Shellman. 2004. Fear of Persecution: Forced Migration 1952–1995. *Journal of Conflict Resolution* 48, 5 (2004), 723–45.
- [47] Brian Neelon, James O'Malley, and Valerie Smith. 2016. Modeling zero-modified count and semicontinuous data in health services research. *Statistics in Medicine* 35, 27 (2016), 5070ff!5093.
- [48] Justin Nix, Bradley A. Campbell, Edward H. Byers, and Geoffrey P. Alpert. 2017. A Bird's Eye View of Civilians Killed by Police in 2015. *Criminology and Public Policy* 16, 1 (2017), 309–340.
- [49] University of Virginia Library. 2016. Getting started with Negative Binomial Regression Modeling. online. (2016). <http://data.library.virginia.edu/getting-started-with-negative-binomial-regression-modeling/> Retrieved Nov 21, 2017.
- [50] George Patterson and Philip Swan. 2016. Police shootings of unarmed African American males: A systematic review. *Journal of human behavior in the social environment* 26, 3-4 (2016), 267–278.
- [51] Yasser Arafat Payne, Brooklynn K. Hitchens, and Darryl L. Chambers. 2017. fiWhy I Can't Stand Out in Front of My House?fi: Street-Identified Black Youth and Young Adult's Negative Encounters With Police. *Sociological Forum* (2017).
- <http://onlinelibrary.wiley.com/doi/10.1111/sofc.12380/full>
- [52] Matthew Petrocelli, Alex RPiquero, and Michael Smith. 2003. Conflict theory and racial profiling: An empirical analysis of police traffic stop data. *Journal of Criminal Justice* 31 (2003), 1–11.
- [53] Washington Post. 2015–2017. Fatal Force. online. (2015–2017). <https://www.washingtonpost.com/graphics/national/police-shootings-2017/>
- [54] William Pridemore. 2005. A Cautionary Note on Using County-Level Crime and Homicide Data. *Homicide Studies* 9, 3 (2005), 256–268.
- [55] Doris Marie Provine. 2011. Race and Inequality in the War on Drugs. *Annual Review of Law and Social Science* 7 (2011), 41–60.
- [56] Victor Rios. 2011. *Punished: Policing the lives of Black and Latino boys*. New York University Press.
- [57] George Ritzer and Jeffery Stepnisky. 2018. *Sociological Theory*. Sage.
- [58] Cathy Lisa Schneider. 2014. *Police Power and Race Riots*. University of Pennsylvania Press, Philadelphia.
- [59] Steven Seidman. 2016. *Contested Knowledge: Social Theory Today*, 6th. Wiley.
- [60] Kuang Keng Kueh Ser. 2016. When it comes to police shootings, the US doesn't look like a developed nation. online. (July 2016). <https://www.pri.org/stories/2016-07-12/when-it-comes-police-shootings-us-doesnt-look-developed-nation> retrieved Nov 21, 207.
- [61] Douglas Sharp and Susie Atherton. 2007. To Serve and Protect?: The Experiences of Policing in the Community of Young People from Black and Other Ethnic Minority Groups. *The British Journal of Criminology* 47, 5 (2007), 746–763.
- [62] Samuel Sinyangwe, DeRay McKesson, and Brittany Packnett. 2013–2017. Mapping Police Violence. online. (2013–2017). <https://mappingpoliceviolence.org/>
- [63] Brad Smith. 2004. Structural and organizational predictors of homicide by police. *Policing: An International Journal* 27, 4 (2004), 539–57.
- [64] Brad W. Smith and Malcolm D. Holmes. 2014. Police Use of Excessive Force in Minority Communities: A Test of the Minority Threat, Place, and Community Accountability Hypotheses. *Social Problems* 61, 1 (2014), 83–104.
- [65] Mikis D. Stasinopoulos, Robert A. Rigby, Gillian Z. Heller, Vlasis Voudouris, and Fernanda De Bastiani. 2017. *Flexible Regression and Smoothing: Using GAMSS in R*. CRC Press.
- [66] Brian Stults and Eric Baumer. 2007. Racial Context and Police Force Size: Evaluating the Empirical Validity of the Minority Threat Perspective. *Amer. J. Sociology* 113, 2 (2007), 507–46.
- [67] Melissa Tetzlaff-Bemiller. 2013. *Child Murder: A Re-examination Of Durkheim's Theory Of Homicide*. Ph.D. Dissertation. University of Central Florida.
- [68] Jeramy Townsley. 2017. GIS Shapefile. online. (2017). <https://iu.app.box.com/file/248794168208>
- [69] Jeramy Townsley. 2017. Original download of Mapping Police Violence data, before cleaning. online. (2017). <https://drive.google.com/drive/folders/1hIUT-165SriYaqzQN5IQXhDp-C7zyZnnK>
- [70] Jeramy Townsley. 2017. Python script and data. online. (2017). <https://github.com/bigdata-i523/hid347/blob/master/project/script/qgisscriptkilledbypolice.py>
- [71] Jeramy Townsley. 2017. R-script Data. online. (2017). <https://github.com/bigdata-i523/hid347/blob/master/project/script/rscript>

[Figure 5 about here.]

[Figure 6 about here.]

## LIST OF FIGURES

1	U.S. county-level map of residents killed by police, 2013-Oct 2017. Data from Census and mappingpoliceviolence.org	11
2	Left: Histogram of the number of county residents killed by police. This histogram shows a maximum of eight per county killed by police, while the actual maximum is 266. However, there are only 120 observations where this value is above 8. This view gives a better depiction of the data. Right: Scatter plot of the total county population versus the total killed by police in that county. The red line is a regression prediction line for these two variables.	12
3	Left: Scatter plot of predicted versus observed counts of residents killed by police per county. Right: Rootogram-fit of the final regression model: negative binomial with three independent variables (log of population, employment and median income). The red line can be interpreted as the observed values. Bars that are hanging below the zero-line are under-fit for that bin, and those above the line are over-fit for that bin.	13
4	Regression diagnostics plots for the final model with three independent variables: population (log), employment and median income. Top: Residuals versus fitted and residuals versus index. Bottom: Quantile residuals and QQ Plot of residuals.	14
5	Portion of the Python QGIS script that downloads the data and maps the county-level killings by United States law enforcement (contiguous counties only)	15
6	Portion of R-script to download the data then performs the regression analysis, diagnostics, and plots.	16

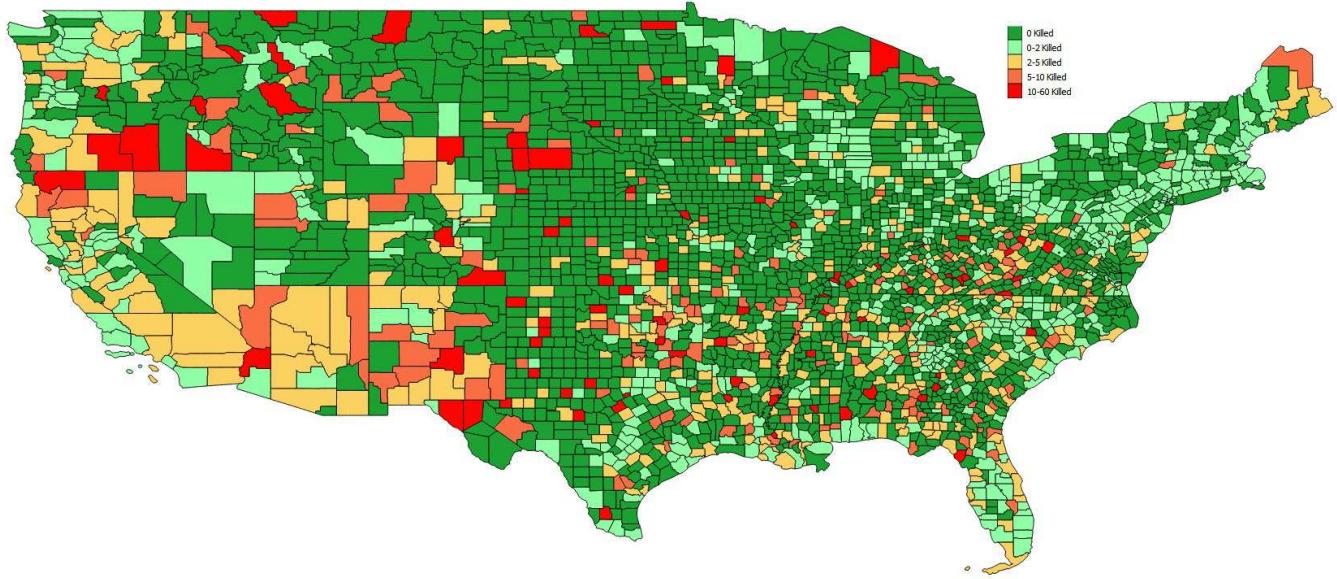


Figure 1: U.S. county-level map of residents killed by police, 2013-Oct 2017. Data from Census and mappingpoliceviolence.org

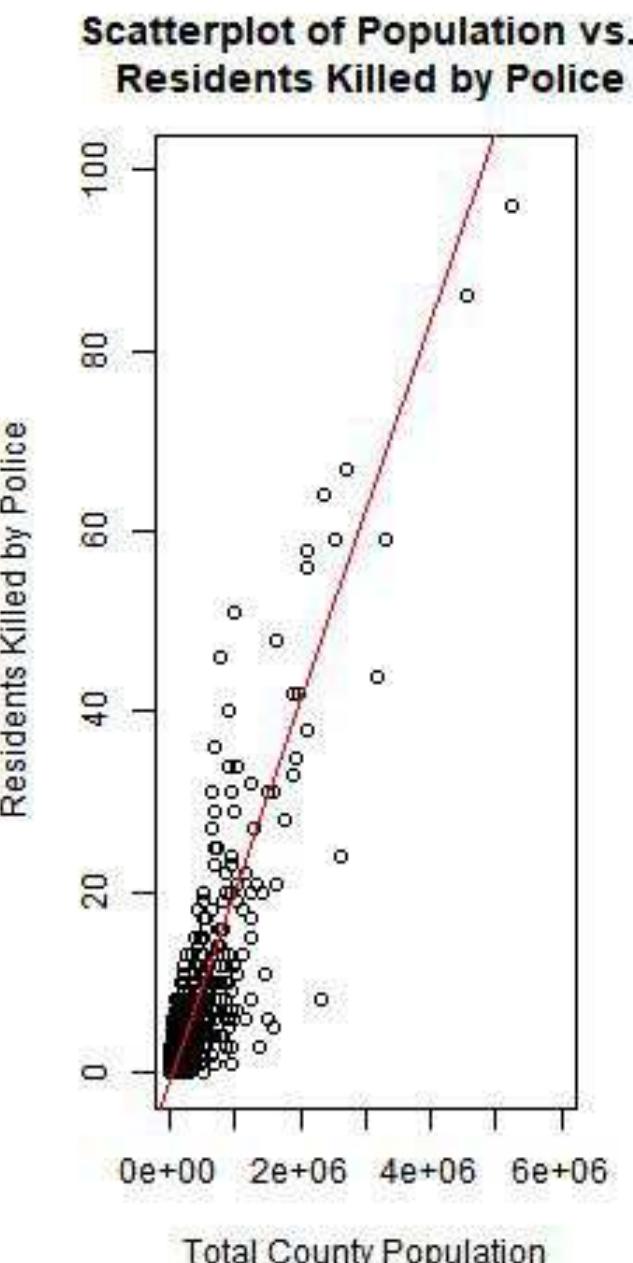
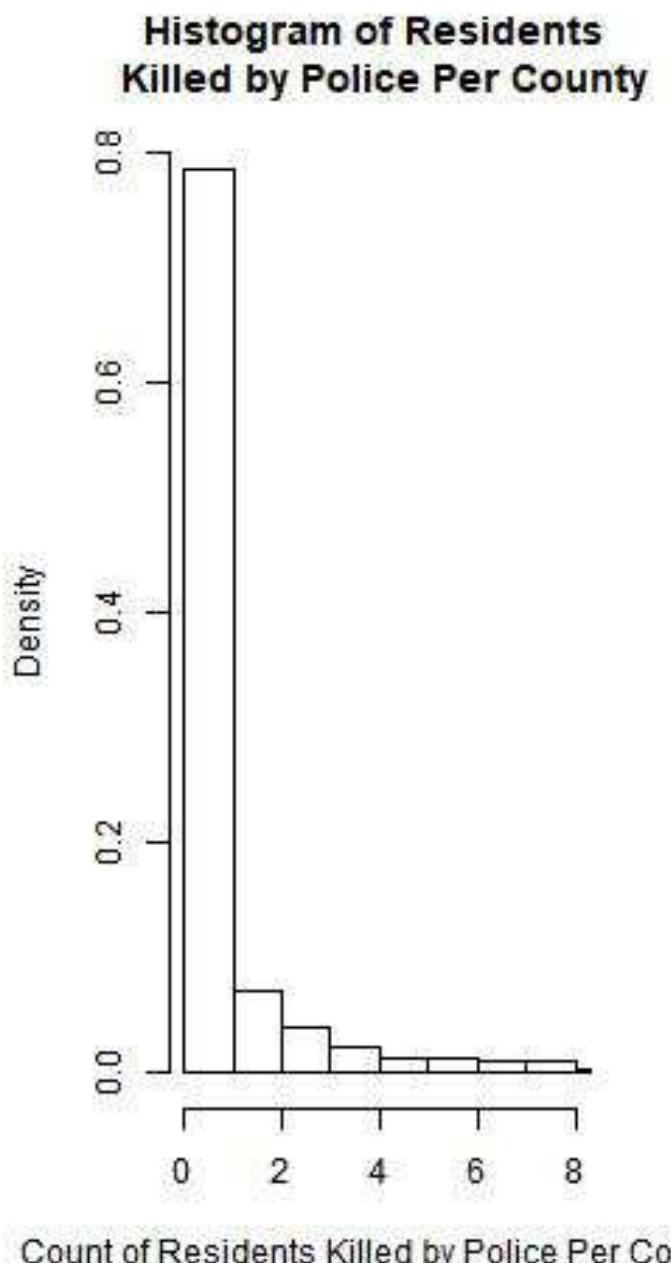


Figure 2: Left: Histogram of the number of county residents killed by police. This histogram shows a maximum of eight per county killed by police, while the actual maximum is 266. However, there are only 120 observations where this value is above 8. This view gives a better depiction of the data. Right: Scatter plot of the total county population versus the total killed by police in that county. The red line is a regression prediction line for these two variables.

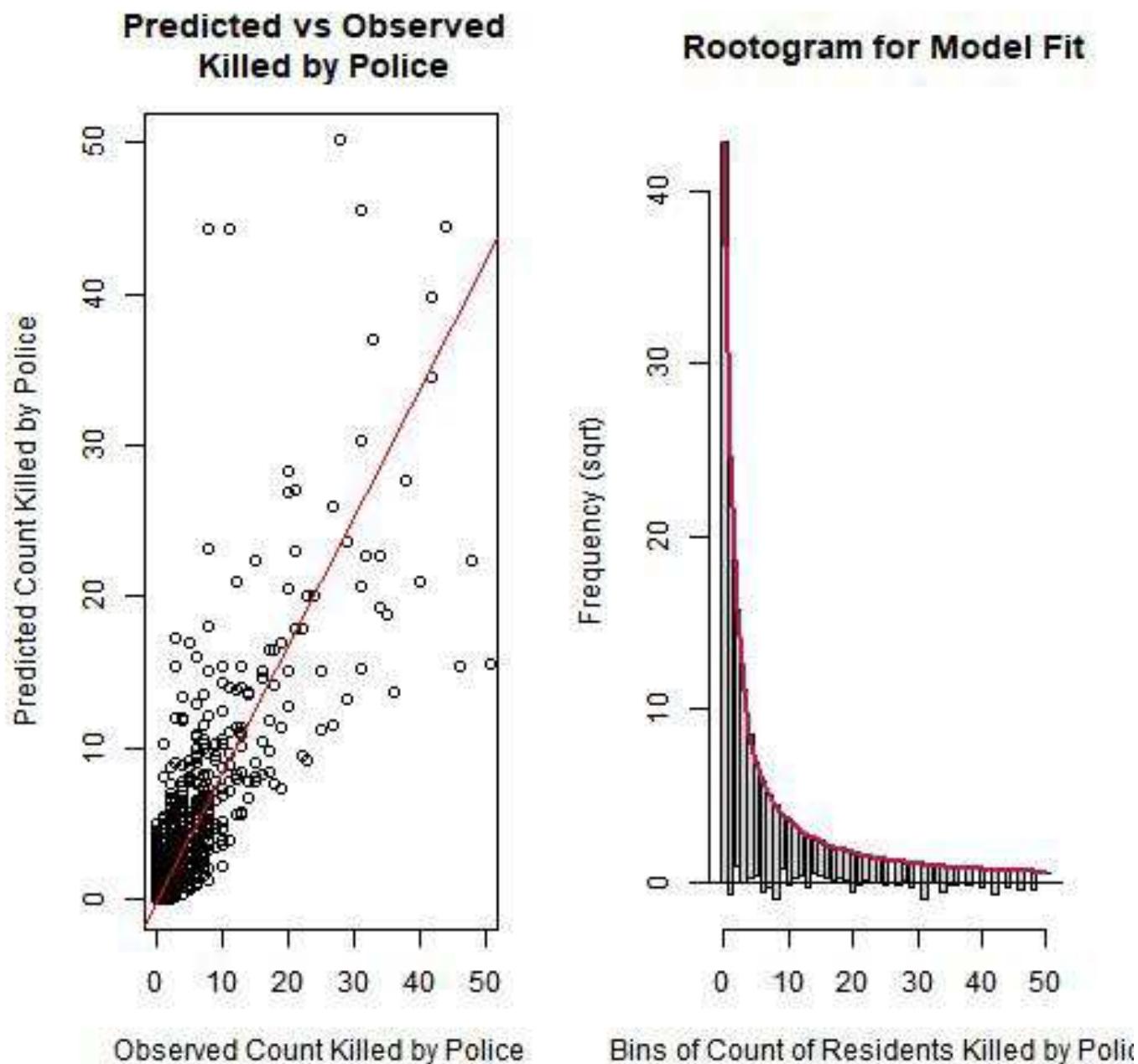


Figure 3: Left: Scatter plot of predicted versus observed counts of residents killed by police per county. Right: Rootogram-fit of the final regression model: negative binomial with three independent variables (log of population, employment and median income). The red line can be interpreted as the observed values. Bars that are hanging below the zero-line are under-fit for that bin, and those above the line are over-fit for that bin.

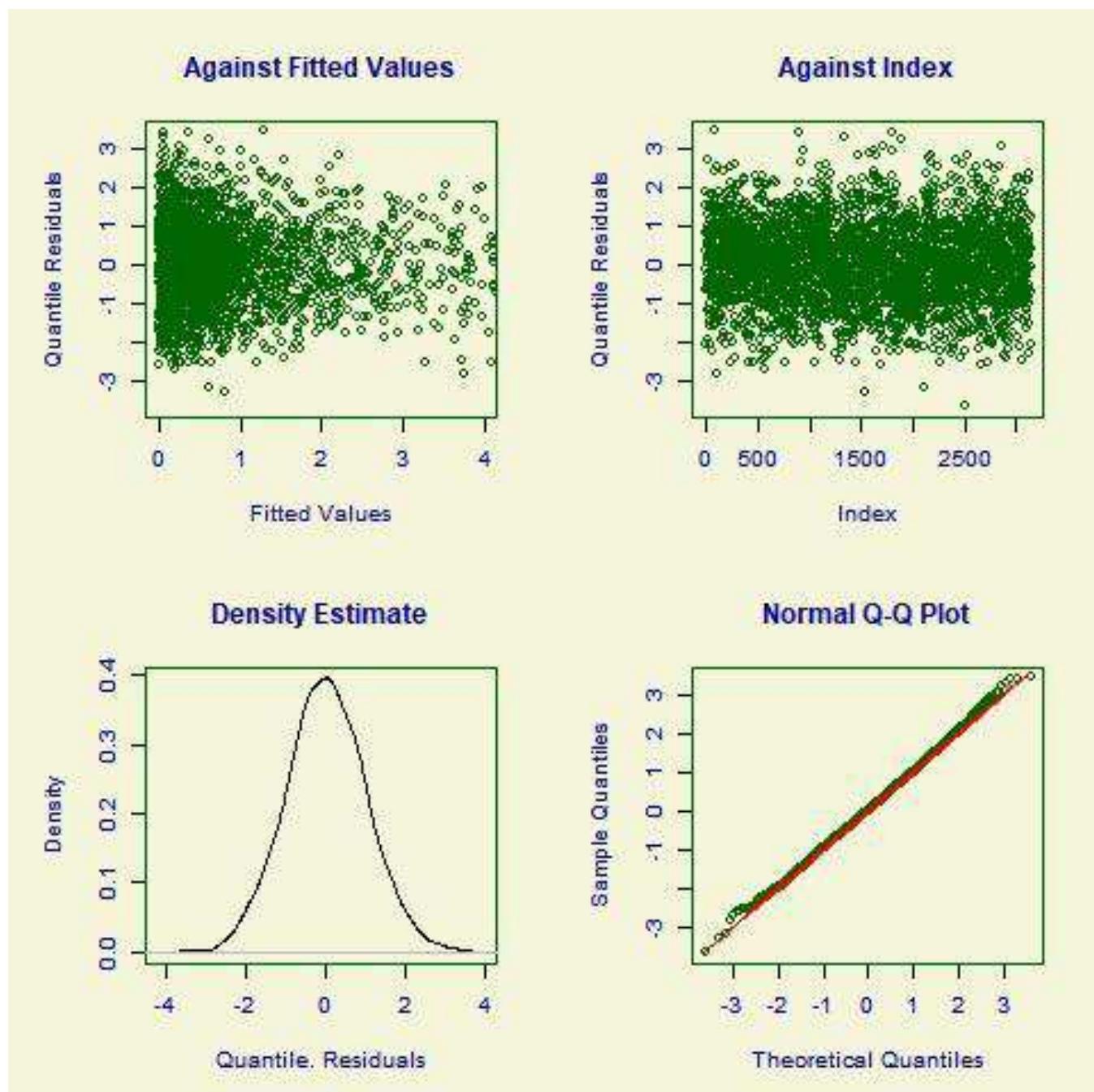


Figure 4: Regression diagnostics plots for the final model with three independent variables: population (log), employment and median income. Top: Residuals versus fitted and residuals versus index. Bottom: Quantile residuals and QQ Plot of residuals.

```

#####
# Python script to download data and perform mapping
# Download shapefile and data
#####
import urllib
urllib.urlretrieve('https://iu.box.com/shared/static/dyd79nfosj25kdqo4hk0me1tmhrbn275.zip', 'counties.zip')
}

#####
# Unzip counties file
#####
from qgis.utils import iface
import zipfile
zip_ref = zipfile.ZipFile('counties.zip', 'r')
zip_ref.extractall('counties')
zip_ref.close()

#####
# Open the shapefile from inside QGIS python plugin
#####
wb=QgsVectorLayer('counties/tl_2017_us_county.shp','counties','ogr')
QgsMapLayerRegistry.instance().addMapLayer(wb)

#####
# Delete everything except contiguous 48 states
#####
features = wb.getFeatures()
deletelist=['02', '15', '60', '66', '69', '72', '78']
ids = [ f.id() for f in features if f.attribute('STATEFP') in deletelist]
with edit( wb ):
    wb.deleteFeatures( ids )

...

```

**Figure 5: Portion of the Python QGIS script that downloads the data and maps the county-level killings by United States law enforcement (contiguous counties only)**

```

#####
# R-script to download data and perform regression analysis
# import libraries
#####
library(gamlss);library(countreg); library(MASS)
library(readr);library(car); library(mgcv);library(RCurl)

#####
# Get data
# demographic variables from American Community Survey (US Census, 5-yr estimate, 2015)
# https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
# Original police violence data from https://mappingpoliceviolence.org/
# police killing data had to be extensively cleaned & corrected (60 bad counties)
#####
getdata <- getURL("https://raw.githubusercontent.com/bigdata-i523/hid347/master/project/script/data1.csv")
data1 <- read.csv(text = getdata)
dt1<-data1[complete.cases(data1),] # Only 2 missing cases
summary(dt1)

#####
# Get 3 models & summaries: Final model (3 IVs), full model
# & (all 10 IVs) & partial model (6 IVs)
#####
finmod<-gamlss(killed ~ log(pop15)+ emp15 + mdhinc15, data=dt1,
                 family=NBI) # removed p>.01, no vif>2
midmod<-gamlss(killed ~ log(pop15)+ emp15 + rural15 + mdhinc15 + wbrat15,
                 data=dt1,family=NBI) # removed p>0.1, no vif>4
fullmod<-gamlss(killed ~ log(pop15)+ emp15 + pov15 + rural15 +
                 mdhinc15+ab100k15 + bagrad15 + nohs2515 + snap15 + wbrat15, data=dt1,family=NBI)
summary(finmod)
summary(midmod)
summary(fullmod)

#####
# Get theta (negative binomial factor), R^2 and Deviance Explained
#####
fintheta<-glm.nb(killed ~ log(pop15)+emp15 +mdhinc15, data=dt1)$theta
fulltheta<-glm.nb(killed ~ log(pop15)+ emp15 + rural15 + mdhinc15 +
                  wbrat15, data=dt1)$theta
fulltheta2<-glm.nb(killed ~ log(pop15)+ emp15 + pov15 + rural15 +
                  mdhinc15+ab100k15 + bagrad15 + nohs2515 + snap15 + wbrat15, data=dt1)$theta
summary(gam(killed~log(pop15)+emp15 +mdhinc15,data=dt1,
            family=negbin(fintheta))) # 0.874, DevExp=76.6%
summary(gam(killed~log(pop15)+ emp15 + rural15 + mdhinc15 + wbrat15,data=dt1,
            family=negbin(fintheta))) # 0.865, DevExp=76.7%
summary(gam(killed~log(pop15)+ emp15 + pov15 + rural15 +
            mdhinc15+ab100k15 + bagrad15 + nohs2515 + snap15 + wbrat15,data=dt1,
            family=negbin(fintheta))) # 0.879, DevExp=76.8%


...

```

**Figure 6: Portion of R-script to download the data then performs the regression analysis, diagnostics, and plots.**

LIST OF TABLES

- |   |  |    |
|---|--|----|
| 1 | All variables used in this study.  | 18 |
| 2 | Regression and diagnostic results from all three models. "Zeros Pred" are the number of counties where no person is predicted to be killed by police, calculated not at exact zero, but any value less than 0.5. "Killed Pred" is the mean number of predicted people killed by police in each county. | 19 |

<b>Variable Name</b>	<b>Description</b>	<b>Min</b>	<b>Max</b>	<b>Mean</b>	<b>SD</b>
killed	Count of people killed by police per county from 2013-Oct 2017	0	266	1.74	7.55
pop15	Total county population	285	10,112,255	10,196	328,221
emp15	Percent of people employed	20.3%	91.3%	69.7%	9.7%
mdhinc15	Median income	\$19,328	\$123,453	\$45,095	\$12,248
rural15	Percent of people living in 'rural' areas	0.0%	100.0%	59.5%	31.5%
wbrat15	White:Black Ratio	0	9,901	140.8	420.5
pov15	Percent of people living in poverty	1.4%	53.3%	16.7%	6.6%
ab100k15	Percent of people making above \$100k per year	1.8%	62.6%	16.1%	7.8%
bagrad15	Percent of people 25 or older with a bachelor's degree or higher	5.0%	78.8%	20.4%	9.0%
nohs2515	Percent of people 25 or older with no high school diploma	1.6%	62.6%	16.1%	6.6%
snap15	Percent of people on SNAP (food stamps)	0.0%	54.1%	14.3%	6.7%
All data (except killed) are from the US Census, American Community Survey, 2015, 5-year estimates, retrieved November 2017.					
Killed is from mappingpoliceviolence.org, with 60 counties needing to be manually corrected through a search of local newspapers where the individual was killed, retrieved November 2017.					

**Table 1: All variables used in this study.**

<b>Measures</b>	<b>Final Model</b>	<b>Partial Model</b>	<b>Full Model</b>
(Intercept)	-10.02***	-9.41***	-9.65***
log(pop15)	1.092***	1.057***	1.054***
emp15	-0.0185***	-0.020***	-0.015***
mdhinc15	-0.000015***	-0.000015***	-0.000024**
rural15		-0.26	-0.25
wbrat15		0.00015	0.00015
pov15			0.011
ab100k15			2.016
bagrad15			-0.0073
nohs2515			0.46
snap15			-0.0091
AIC	6900	6898	6898
RMSE	2.67	2.77	2.61
Cor Pred	0.936	0.934	0.938
Max VIF	1.99	3.37	34.5
Adjusted R-Sqr	0.874	0.865	0.879
Deviance Expl	76.6%	76.7%	76.8%
Zeros Pred	1656	1670	1668
Killed Pred	1.775	1.619	1.787
<b>Observed</b>			
			1827
			1.739
·p ≤ 0.10; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001			

Table 2: Regression and diagnostic results from all three models. "Zeros Pred" are the number of counties where no person is predicted to be killed by police, calculated not at exact zero, but any value less than 0.5. "Killed Pred" is the mean number of predicted people killed by police in each county.

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--no key, author in qgis  
Warning--no author, editor, organization, or key in qgis  
Warning--to sort, need author or key in qgis  
Warning--no key, author in python  
Warning--no author, editor, organization, or key in python  
Warning--to sort, need author or key in python  
Warning--no key, author in pyqgis  
Warning--no author, editor, organization, or key in pyqgis  
Warning--to sort, need author or key in pyqgis  
Warning--no key, author in r  
Warning--no author, editor, organization, or key in r  
Warning--to sort, need author or key in r  
Warning--no key, author in rstudio  
Warning--no author, editor, organization, or key in rstudio  
Warning--to sort, need author or key in rstudio  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3085 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Warning--no key, author in pyqgis  
Warning--no key, author in pyqgis  
Warning--no key, author in python  
Warning--no key, author in python  
Warning--no key, author in qgis  
Warning--no key, author in qgis  
Warning--no key, author in r

```
Warning--no key, author in r
Warning--no key, author in rstudio
Warning--no key, author in rstudio
Warning--no key, author in pyqgis
Warning--no author, editor, organization, or key in pyqgis
Warning--empty author in pyqgis
Warning--no key, author in python
Warning--no author, editor, organization, or key in python
Warning--empty author in python
Warning--no key, author in qgis
Warning--no author, editor, organization, or key in qgis
Warning--empty author in qgis
Warning--no key, author in r
Warning--no author, editor, organization, or key in r
Warning--empty author in r
Warning--no key, author in rstudio
Warning--no author, editor, organization, or key in rstudio
Warning--empty author in rstudio
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry beaujean17
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry beaujean17
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry beaujean17
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry beaujean17
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--no number and no volume in currie16
Warning--page numbers missing in both pages and numpages fields in dalton17
Warning--can't use both author and editor fields in feng16
Warning--empty address in feng16
Warning--empty address in foucault77
Warning--empty address in fox15
Warning--empty address in jones11
Warning--empty address in luders10
Warning--no number and no volume in payne17
Warning--page numbers missing in both pages and numpages fields in payne17
Warning--empty address in rios11
Warning--empty address in ritzer18
Warning--empty address in seidman16
Warning--empty address in gammelss
(There were 12 error messages)
make[2]: *** [bibtex] Error 2
```

latex report

[2017-12-16 09.39.27] pdflatex report.tex

..../README.yml

11:81	error	line too long (1004 > 80 characters)	(line-length)
17:27	error	trailing spaces (trailing-spaces)	
29:81	error	line too long (1004 > 80 characters)	(line-length)
35:12	error	trailing spaces (trailing-spaces)	
42:81	error	line too long (116 > 80 characters)	(line-length)
43:81	error	line too long (114 > 80 characters)	(line-length)
43:114	error	trailing spaces (trailing-spaces)	
44:81	error	line too long (112 > 80 characters)	(line-length)
44:112	error	trailing spaces (trailing-spaces)	
45:81	error	line too long (116 > 80 characters)	(line-length)
45:116	error	trailing spaces (trailing-spaces)	
46:81	error	line too long (118 > 80 characters)	(line-length)
46:118	error	trailing spaces (trailing-spaces)	
47:81	error	line too long (118 > 80 characters)	(line-length)
47:118	error	trailing spaces (trailing-spaces)	
48:81	error	line too long (113 > 80 characters)	(line-length)
48:113	error	trailing spaces (trailing-spaces)	
49:81	error	line too long (116 > 80 characters)	(line-length)
49:116	error	trailing spaces (trailing-spaces)	

## Compliance Report

```
name: Jeramy Townsley  
hid: 347  
paper1: Oct 21 17 100%  
paper2: Nov 05 17 100%  
project: Nov 29 17 100%
```

yamlcheck

---

wordcount

---

19

```
wc 347 project 19 8681 report.tex
wc 347 project 19 10200 report.pdf
wc 347 project 19 2515 report.bib
```

find "

---

103: This data contained errors that had to be manually corrected.

The errors were discovered when the "killed" variable, count of the number of people killed by law enforcement in any given county, was merged with Census data, the independent variables. The Census data contains an official list of counties in every state. The MPV data listed counties that were not in the claimed states. It was discovered that some were simply errors, some were cities coded as counties, and some counties were blank. Sixty observations total were found to have errors. Using the victim's name provided by MPV, they were searched using Google (November, 2017), to find an original news story reporting this killing. The pre-cleaned data is available on the author's Google Drive course site. \cite{townsleyG} Once the killing could be verified, similar investigation confirmed the county and state of the killing. The corrected data was used for this analysis, and is available from the author's Github course site.

\cite{townsleyR} Every missing or incorrect county was found. It is possible that other errors exist, since the only errors detected were those where there was no correct match with an actual Census-listed county and state. There may be counties where a victim is claimed to have been killed, where the county and state exist, but misidentified. The data has approximately the same number of observations as the two news agencies' data, strengthening its claims to validity. Further, each of the 60 misspecified counties that were corrected, traced back to an actual killing by law enforcement based on a search of the victim's name, increasing confidence in the data.

127: Since the data used here is coded into a county-level datafile, the first entry in the dbf is Cuming County, Nebraska, with Nebraska geocoded as state \#31, and Cuming County geocoded as

county \#039. All 3,142 counties in the dbf file have these coded identifiers that are matched to demographic and employment data that can be downloaded separately from the Census Bureau, such as ACS data. Since the geocodes are identical across all data sets for each specific geography, this data and the shapefiles can be readily integrated. Integrating other types of data, such as the police killings data from MPV, requires a different approach. In this case, since MPV provided both the county and state names where the deaths occurred, merging these into one string, such as "CumingNE", and doing the same with the TIGER shapefile data, allows the two files to be jointly identified and merged. For Figure 1, which shows the rates of killings by police at the county-level, a ratio was calculated in QGIS--the count of those killed in each county divided by the total population of that county for the measure, rates people killed by police per 100,000. With the creation of this new variable, it can be mapped in QGIS as a {\em graduated symbol} with a pre-specified color scheme.

164: \caption{Regression and diagnostic results from all three models. "Zeros Pred" are the number of counties where no person is predicted to be killed by police, calculated not at exact zero, but any value less than 0.5. "Killed Pred" is the mean number of predicted people killed by police in each county.}

254: getdata <- getURL("https://raw.githubusercontent.com/bigdata-i523/hid347/master/project/script/data1.csv")

passed: False

find footnote

---

16: \renewcommand\footnotetextcopyrightpermission[1]{} % removes footnote with conference information in first column

passed: False

find input{format/i523}

---

passed: False

find input{format/final}

---

passed: False

floats

---

108: \begin{table}  
109: \includegraphics[width=1.0\textwidth]{images/table1.jpg}  
129: \begin{figure}  
130: \includegraphics[width=1.0\textwidth]{images/figure1.jpg}  
134: This process can be automated with Python, a programming language useful for big data analysis. \cite{python} QGIS has integrated Python into its software with the inclusion of a Python Console for direct use, and an online instruction manual available through the QGIS web site. \cite{pyqgis} For the map in Figure 1, a Python script was created that downloads a zipped file containing the TIGER shapefile for the US counties, and all of the census data that had previously been integrated into the dbf file. The script and data are available on the author's github course site. \cite{townsleyP} The zipped shapefile is available on the author's Indiana University Box account (it was too large for Github to allow). \cite{townsley2} The Python program unzips the file, deletes all parts of the map except for the contiguous United States (all but 48 states), then applies a graduated color scheme to the killed per population variable. Figure \ref{townsleyP} shows part of the code to download and extract the data.  
137: Regression analysis can be performed, and graphs created, through open source software, R, an environment for statistical analysis and graphics. \cite{r} In addition to the base R package, which is command line only, other packages are available as an overlay to incorporate GUI features, such as RStudio. \cite{rstudio} This analysis was done with R 3.4.1, using RStudio 1.0.153. Instead of Python, R has its own scripting language. The R-script for the following procedures and data are available on the author's github course site. \cite{townsleyR} Figure \ref{townsleyR} shows the code to download and extract the data.  
143: \begin{figure}  
144: \includegraphics[width=1.0\textwidth]{images/figure2.jpg}  
162: \begin{table}  
163: \includegraphics[width=1.0\textwidth]{images/table2.jpg}  
177: \begin{figure}  
178: \includegraphics[width=1.0\textwidth]{images/figure3.jpg}  
184: \begin{figure}  
185: \includegraphics[width=1.0\textwidth]{images/figure4.jpg}  
199: \begin{figure}[htb]  
235: \caption{Portion of the Python QGIS script that downloads the

```
data and maps the county-level killings by United States law
enforcement (contiguous counties only)}\label{townsleyP}
238: \begin{figure}[htb]
292: \caption{Portion of R-script to download the data then performs
the regression analysis, diagnostics, and plots.
}\label{townsleyR}
```

```
figures 6
tables 2
includegraphics 6
labels 2
refs 2
floats 8
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are referred to: (refs >= labels)
```

#### Label/ref check

105: Summary statistics and definitions can be found in Table 1 for all of the variables used. The independent variables were all obtained from the United States Census, which maintains a large, publicly-available database on their website. \cite{census} All of the data used for this study are at the administrative unit of the county, and includes all of the 3,142 United States counties listed by the Census. All of the ten variables were downloaded (November, 2017) from the American Community Survey (ACS). The ACS is a subsidiary of the Census Bureau which administers an extensive survey of an annual sampling of the U.S. population on various economic and demographic questions. Each of the variables chosen for this project falls into one of five categories: population total, race, education, geography-type and economic. The total population of each county is estimated by the ACS based on decennial census data, and represents a single-year value. The rest of the variables are estimated from aggregates of data collected from 2011-2015, a five-year estimate. Multiple studies have shown that for homicide analysis, whether resident-on-resident killings, killings by police, or killings of police, demographic and economic variables tend to have strong predictive value.

\cite{pridemore05,kaminski05,legewie15,patterson16,smith14}

117: In addition to demographic and economic data, the Census Bureau provides a significant amount of geospatial information. Through their TIGER products (Topologically Integrated Geographic Encoding and Referencing), they provide a web site where the

public can download shapefiles at many different levels that are published each year, and can be connected to the rest of their data. \cite{tiger} For this study, county-level data was used, since it was the smallest administrative unit for which all of the data was available. While accurate state-level data was available, it did not seem fine-grained enough to provide a sufficient analysis. Smaller administrative levels, like census-tract, and even block-level shapefiles are available, and there is annual census data that can be mapped onto those levels.

While some ACS variables are available at these small units, they typically only cover a random selection, and can have high error rates. County-level data is the smallest administrative unit for which every county in the U.S. can be measured in the five-year estimates file. The shapefile used for this study is the 2015 county-level data for the entire United States, and was downloaded from the Census TIGER site (November, 2017).

\cite{tiger} However, not all places available on the shapefile have data through the ACS. For example, several of the smaller island territories are not included in the ACS survey, so those are excluded in the map, as well as the regression analysis.

While the latter includes Alaska and Hawaii, for ease of viewing the county-level data, those two states are not included on the map (Figure 1).

- 126: Since the data used here is coded into a county-level datafile, the first entry in the dbf is Cuming County, Nebraska, with Nebraska geocoded as state \#31, and Cuming County geocoded as county \#039. All 3,142 counties in the dbf file have these coded identifiers that are matched to demographic and employment data that can be downloaded separately from the Census Bureau, such as ACS data. Since the geocodes are identical across all data sets for each specific geography, this data and the shapefiles can be readily integrated. Integrating other types of data, such as the police killings data from MPV, requires a different approach. In this case, since MPV provided both the county and state names where the deaths occurred, merging these into one string, such as "CumingNE", and doing the same with the TIGER shapefile data, allows the two files to be jointly identified and merged. For Figure 1, which shows the rates of killings by police at the county-level, a ratio was calculated in QGIS--the count of those killed in each county divided by the total population of that county for the measure, rates people killed by police per 100,000. With the creation of this new variable, it can be mapped in QGIS as a {\em graduated symbol} with a pre-specified color scheme.

- 133: This process can be automated with Python, a programming language useful for big data analysis. \cite{python} QGIS has integrated

Python into its software with the inclusion of a Python Console for direct use, and an online instruction manual available through the QGIS web site. \cite{pyqgis} For the map in Figure 1, a Python script was created that downloads a zipped file containing the TIGER shapefile for the US counties, and all of the census data that had previously been integrated into the dbf file. The script and data are available on the author's github course site. \cite{townsleyP} The zipped shapefile is available on the author's Indiana University Box account (it was too large for Github to allow). \cite{townsley2} The Python program unzips the file, deletes all parts of the map except for the contiguous United States (all but 48 states), then applies a graduated color scheme to the killed per population variable. Figure \ref{townsleyP} shows part of the code to download and extract the data.

- 140: Count data with a mass at zeros can be analyzed in a number of ways, depending on the theoretical reason for the zeros, each of which requires a different analytic approach. One decision factor includes knowing whether there is one process generating all of your data, or whether there are two--one generating the count data (above zero) and some of the zero, and a second process generating {\em an excess} of zeros.  
\cite{neelon16,farewell17,min02,martin17} In the latter case, these zeros would be more than would be predicted from the given distribution, and those are sometimes called {\em inaccessible} zeros. Inaccessibility zeros are considered excessive, and a two-part, zero-inflated model might be a good option in that case. Zero-inflated Poisson, and zero-inflated negative binomial models are available in R, in packages such as {\em gamlss} and {\em pscl}. \cite{gamlss} In Figure 2, the {\em killings by police per county} dependent variable is shown in a histogram, and as a plot against the county population, both graphs being created in R using the standard {\em hist} and {\em plot} functions. This allows one to visualize the large mass at zero, which might lead one to presume a zero-inflated model is most appropriate.
- 152: One of the assumptions of a Poisson distribution is that the mean and dispersion are equivalent. A model with a variance greater than the mean is considered over-dispersed, which implies it is a better candidate for a negative binomial regression approach.  
\cite{fox15,beaujean16,smith14,legewie15} As can be seen in Table 1, the mean of the count killed by police is 1.74, and the variance is 57.0 (square of standard deviation, 7.55). This data is clearly over-dispersed, and is thus a good candidate for negative binomial regression.
- 159: Three regression models were created: the full model, which

includes all ten independent variables; the partial model, which includes only five independent variables; and the final model, which includes only the three predictors with the strongest p-values. The coefficients and p-values of these models can be found in Table 2, along with a number of goodness-of-fit and diagnostic information. From the full model, all of the non-significant variables were removed to generate the partial model. From the partial model, the two variables whose p-values were greater than 0.05 were removed from the model, leaving just three predictor variables, all of which were significant to  $p \leq 0.001$ . `\cite{beaujean16}` By all measures, the three models have very similar model fit and diagnostic criteria (AIC, R-square, correlation between predicted and observed, RMSE, predicted mean of outcome variable, etc.). The variance inflation factor was checked for each, and while significant problems were detected for the full model (max VIF=34.5), those problems were eliminated in both of the smaller models. The final model is arguably the best model since it is the simplest, with only three predictor variables, and having equivalent goodness-of-fit and diagnostic values as the more complex models.

- 181: In addition to goodness of fit, regression also requires that residuals meet basic assumptions. Figure 4 shows several standard diagnostic plots indicating that these assumptions seem to be met. The residuals versus fitted values plot is interpreted as being problematic if a pattern emerges from the residuals, either above or below the zero-line. In this case, no pattern is apparent--the residuals seem equally and randomly scattered above and below the line. Similarly, the plot of residuals versus index shows a similar result, indicating the regression residuals assumption is met. Finally, the bottom two plots, more residuals plots, show the same. For example, the Normal Q-Q Plot would indicate a problem if a significant number of dots were straying from the main central line--in this case they are almost all on the line. `\cite{hocking}`

passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

- 109: `\includegraphics[width=1.0\textwidth]{images/table1.jpg}`

130: \includegraphics[width=1.0\textwidth]{images/figure1.jpg}

144: \includegraphics[width=1.0\textwidth]{images/figure2.jpg}

163: \includegraphics[width=1.0\textwidth]{images/table2.jpg}

178: \includegraphics[width=1.0\textwidth]{images/figure3.jpg}

185: \includegraphics[width=1.0\textwidth]{images/figure4.jpg}

passed: False

below\_check

---

WARNING: table and below may be used improperly

175: Figures 3 shows two graphs supporting the conclusion that the final model is an acceptable fit. The left plot of predicted versus observed killings by police per county, with the red regression line of just these two variables, shows the strong relationship between these two variables. The rootogram on the right shows that few bins of predicted outcomes of the final model are significantly different from the observed values. A rootogram is interpreted by the {\em hanging} bars--bars hanging above the zero-line are over-fit, while bars hanging below the line are under-fit. \cite{rootogram} The closeness of the bars to the zero-line indicate a relatively good fit.

WARNING: table and above may be used improperly

175: Figures 3 shows two graphs supporting the conclusion that the final model is an acceptable fit. The left plot of predicted versus observed killings by police per county, with the red regression line of just these two variables, shows the strong relationship between these two variables. The rootogram on the right shows that few bins of predicted outcomes of the final model are significantly different from the observed values. A rootogram is interpreted by the {\em hanging} bars--bars hanging above the zero-line are over-fit, while bars hanging below the line are under-fit. \cite{rootogram} The closeness of the bars to the zero-line indicate a relatively good fit.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no key, author in qgis
Warning--no author, editor, organization, or key in qgis
Warning--to sort, need author or key in qgis
Warning--no key, author in python
Warning--no author, editor, organization, or key in python
Warning--to sort, need author or key in python
Warning--no key, author in pyqgis
Warning--no author, editor, organization, or key in pyqgis
Warning--to sort, need author or key in pyqgis
Warning--no key, author in r
Warning--no author, editor, organization, or key in r
Warning--to sort, need author or key in r
Warning--no key, author in rstudio
Warning--no author, editor, organization, or key in rstudio
Warning--to sort, need author or key in rstudio
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3131 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3131 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3131 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3131 of file ACM-Reference-Format.bst
Warning--no key, author in pyqgis
Warning--no key, author in pyqgis
Warning--no key, author in python
Warning--no key, author in python
```

```
Warning--no key, author in qgis
Warning--no key, author in qgis
Warning--no key, author in r
Warning--no key, author in r
Warning--no key, author in rstudio
Warning--no key, author in rstudio
Warning--no key, author in pyqgis
Warning--no author, editor, organization, or key in pyqgis
Warning--empty author in pyqgis
Warning--no key, author in python
Warning--no author, editor, organization, or key in python
Warning--empty author in python
Warning--no key, author in qgis
Warning--no author, editor, organization, or key in qgis
Warning--empty author in qgis
Warning--no key, author in r
Warning--no author, editor, organization, or key in r
Warning--empty author in r
Warning--no key, author in rstudio
Warning--no author, editor, organization, or key in rstudio
Warning--empty author in rstudio
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Alexander Beaujean and Grant B. Morgan," has a comma at the end for entry bea
while executing---line 3229 of file ACM-Reference-Format.bst
Warning--no number and no volume in currie16
Warning--page numbers missing in both pages and numpages fields in dalton17
Warning--can't use both author and editor fields in feng16
Warning--empty address in feng16
Warning--empty address in foucault77
Warning--empty address in fox15
Warning--empty address in jones11
Warning--empty address in luders10
Warning--no number and no volume in payne17
Warning--page numbers missing in both pages and numpages fields in payne17
Warning--empty address in rios11
Warning--empty address in ritzer18
Warning--empty address in seidman16
Warning--empty address in gamlss
(There were 12 error messages)
```

bibtex\_empty\_fields

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# Analysis of Digit Recognizer classification algorithms in big data

Junjie Lu

Indiana University Bloomington  
3322 John Hinkle Place  
Bloomington, Indiana 47408  
junjlu@iu.edu

Yuchen Liu

Indiana University Bloomington  
1750 N Range Rd  
Bloomington, Indiana 47408  
liu477@iu.edu

Wenxuan Han

Indiana University Bloomington  
1150 S Clarizz Blvd  
Bloomington, Indiana 47401-4294  
wenxhan@iu.edu

## ABSTRACT

Digit Recognizer is becoming more and more important in many different areas, such as zip code recognizer, banking receipt and balance sheet. Many technology companies are trying to use Big Data to develop more efficient and accurate algorithm for Digit Recognizer. This project uses Digit Recognizer data set from Kaggle.com. There are more than 42000 samples in the data set. Each sample contains 784 features which contain pixel information from a  $28 \times 28$  graph. Each pixel has a value between 0 to 255. We use binary classification technique for data cleaning and PCA for feature extraction. For the classification model, we choose five most commonly used classification algorithms, which include Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). From the result, SVM classifier on PCA data produces the highest accuracy with 0.9813. The time spend is 127 seconds. Naive Bayes classifier on PCA data spends the least amount of time to finish the classification task. It takes less one second and reaches a 0.8651 accuracy.

## KEYWORDS

I523, HID213, HID214, HID209, Big Data, Digit Recognition, Cross Validation, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine

## 1 INTRODUCTION

People have made a great improvement in digital recognition in recent years. And it plays significant roles in many different areas. Zip code recognizer can scan zip code for post office automatically. Recognizer in banks can help managing user account by scanning their account number. They help people a lot in increasing working efficiency. And many new productions use digital recognition to authenticate password. In this situation, the accuracy and efficiency of recognition become more and more essential and methods in order to increase the accuracy and efficiency are also required.

Fortunately, people have already developed many different types of techniques to avoid faults and decrease running time in recent few years. Several algorithms will be mentioned here. Logistic regression, the most frequently used algorithm in the field of machine learning, also has a good performance in digital recognition. Decision tree is commonly used in decision analysis. It can identify strategies to get a result, in this case, it can also play an important role in digital recognition. Naive Bayes classifier would also be used. Random forest is also a widely used technique in the field of classification and regression. Its special structure with the multitude of decision trees would help it get a fantastic result. Support vector machine can efficiently perform non-linear classification

hence it also be considered frequently. These algorithms have different structures so that they have different performance. We can also observe running time and accuracy of different algorithms with different kind of data. In this paper, we are going to talk about this and make the comparison between algorithms in accuracy and efficiency.

## 2 DIGIT RECOGNITION APPLICATION IN BIG DATA

Digit recognition have many applications in our daily life. More and more organizations start using Digit recognition technique to save cost and increase accuracy in large-scale data entry jobs.

First, Digit recognition can be used in large scale statistic. For example, it can be used in industry annual inspection and population census [4]. The United States Census Bureau will lead a population census every year in United States, the data volume is huge. A large amount of data needs to be input. In the past, all the data need to input into the database manually. This process required large amount labor force and human resources. In recent years, more and more companies and countries start to use OCR and Digit recognition technique for these jobs. Because the dataset from these applications are centrally organized. Usually, we can build forms automatically and impose restrictions on writing to facilitate automatic Digit Recognition. At present, most of these applications required user to fill in the assigned boxed according to specific requirements. Also, in order to check the accuracy of the recognition, these systems tend to use a user interface to make a comprehensive examination of the recognition results [8]. Currently, more and more advanced algorithms are used in Digit recognition.

Second, Digit recognizer has a widely used in finance and tax administration. As the development of the economy in the world, there are more and more reports, forms, checks and bills waiting to be dealt. The person may deal with these in a comparable lower efficient. It is fantastic for the appearance of digit recognizer to work with these [24]. It has a higher efficiency and longer working hours. And machines do not need the salary. It is more difficult in recognize checks and forms because of higher demand for accuracy. In the meanwhile, there may be several kinds of forms. Recognizer should have the capability to deal with them all. Furthermore, recognizer has to face millions of handwriting some of them are hard to recognize [3].

Third, this technology is also popularly applied in mail sorting. With the improvement of people's living standards, the development of economic activities, the demand for communication causes a substantial increase in the exchange of letters. For example, the post offices in metropolitan areas of China have already reached

to a few million pieces per day in 2000. This sharp rise in business volume has made the sorting of mail pieces automated and becoming the trend of the times. In the automatic sorting of mail, handwritten digit recognition (OCR) often combined with optical bar code recognition (OBR) and artificial identification to complete the postal code reading. Currently, the most common used OVCS sorter has the performance with 30% OCR rejection rate and 1.1% OCR sorting error rate [21].

### 3 EXPERIMENT PREPARATION

In this paper, we choose the data of Digit Recognizer from Kaggle.com in order to test different classification algorithms [7]. The goal of this experiment is to correctly identify digits from a data set of tens of thousands of handwritten images. Thus, we could compare the pros and cons of each technique through the recognition accuracy and time-consuming.

#### 3.1 Data Set Description

In train.csv data file, it contains 42000 gray-scale images of hand-drawn digits, from zero through nine. Each image is a  $28 \times 28$  pixels matrix with a total of 784 pixels [7]. Each pixel has a single pixel-value which is an integer from 0 to 255 associated with it, indicating the lightness or darkness of that pixel (higher numbers meaning lighter). In this experiment, we have plotted the graph in order to see the appearance of these digits easily. Figure 1 shows the first 70 samples.

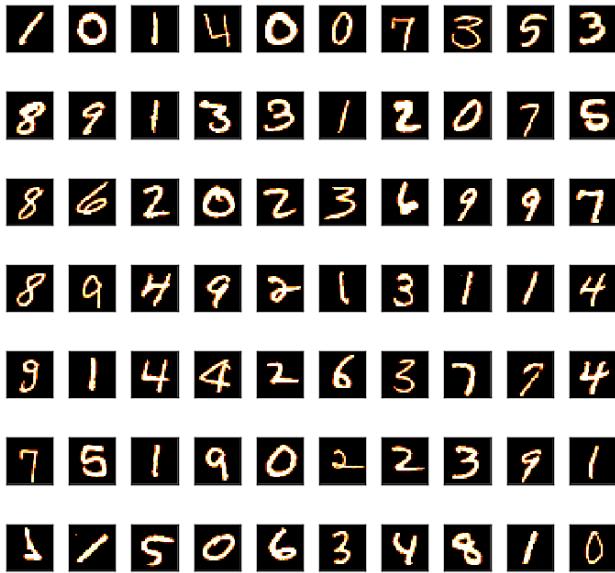


Figure 1: 70 samples of hand-drawn digits in this data set

The training data set has 785 columns. The first column called “label”, is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. Each pixel column in the training set has a name like  $pixelx$ , where  $x$

is an integer between 0 and 783. To locate a pixel on the image, suppose that we have decomposed  $x$  as  $x = i * 28 + j$ , where  $i$  and  $j$  are integers between 0 and 27. Then  $pixelx$  is located in row  $i$  and column  $j$  of this matrix [7]. Visually, if we omit the “pixel” prefix, the pixels make up the image like the following form:

000	001	002	003	...	026	027
028	029	030	031	...	054	055
056	057	058	059	...	082	083
:	:	:	:	:	:	:
728	729	730	731	...	754	755
756	757	758	759	...	782	783

#### 3.2 Data Cleaning

As we mentioned above, it can be seen from both the figure and the pixel-value that the value varies from 0 to 255, which means each feature is a continuous value. Thus, it is possible that such continuous values might affect our later feature selection. Our observation shows that the values are not very high at the boundaries of 0 and  $> 0$ . So here exist three ways to handle it [28]:

- (1) Not do any processing on image;
- (2) Binarize the image. That is, for the values which are 0, keep them as 0; for the values which are greater than 0, change to 1;
- (3) Binarize the image by setting a threshold. That is, for the values which are greater than this threshold, change to 1; otherwise, change to 0.

Obviously, method (2) and (3) will cause the loss of the original information. However, this information may not as important as our expected during the execution of classification algorithms, it could play a positive role in increasing the performance without reducing the accuracy.

In our experiment, we selected method (2) to clean the raw data. The main codes in Figure 2 shows this operation.

```
from numpy import *
# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0
    print("Data clean completed.")
    return new_data
```

Figure 2: The core codes about data clean

### 3.3 Feature Extraction

Dimension reduction in the field of machine learning refers to using a mapping method to map the data points in the original high-dimensional space into the low-dimensional space. The essence of dimension reduction is to learn a mapping function  $f : x \rightarrow y$ , where  $x$  is the expression of the original data point,  $y$  is the low-dimensional vector representation after the data point mapping [12].

The reason why we use data after dimension reduction is that the redundant information and noise information are contained in the original high-dimensional space, which reduces the accuracy of our model. By dimension reduction, we hope to reduce the error caused by redundant information and improve the accuracy of identification. We also hope to find the intrinsic structure of the data structure through the dimension reduction algorithm. Also, in this example, there are 784 features in our data. Space, time and computation complexity are all unacceptable. There are many different dimension reduction algorithms for us to choose. In this project, we choose to use Principle Component Analysis (PCA).

#### 3.3.1 PCA

Principal Component Analysis (PCA) is the most commonly used method of supervised linear dimension reduction. Its goal is to map high-dimensional data to a low-dimensional representation of space by some kind of linear projection. The variance of the data is expected to be maximized in the projected dimension. By keep the variance of data as high as possible, PCA can reduce the dimension of data and keep the loss of information of the data as a minimum [5].

A common understanding is that if all the points are mapped together, almost all information (such as the distance between points) is lost. If the post-mapping variance is as large as possible, the data points are spread apart to preserve more information. It can be proved that PCA is a linear dimension reduction method that loses the original least data information.

One of the questions we faced while we are using PCA is that: how many components should we choose for the model after dimension reduction. In order to solve this problem, we use Explained Variance as our threshold standard. Explained Variance is an important indicator of PCA dimension reduction. The Explained Variance shows the amount of variance explained by each of the selected components. The first column of the PCA model always explains the most variance and the variance explained will keep decrease as the number of column increase. Generally, a dimension with a cumulative contribution rate of about 90% is selected as a reference dimension for PCA dimensionality reduction. In this project, in order to get a more accurate result, we choose 95% as our threshold.

After calculating the explained variance for each component, we decide to choose 30 components for our model. Which shows that there will be 30 features in our model.

## 4 EXPERIMENT ALGORITHMS

We aim to select five most commonly used classification algorithms which include Decision Tree, Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). This section offers a broad overview of these algorithms before applying

```
from sklearn.decomposition import PCA

def feature_selection(data):
    pca = PCA()
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            break

    # Then, PCA the model by the number of components
    pca = PCA(n_components=n, whiten=True)
    return pca.fit_transform(data)
```

Figure 3: The core codes about PCA processing

them to the digit recognizer problem to compare their characteristics. Then, the result of the different algorithm on different data will show on a table.

For each algorithm, we use:

- (1) PCA data - data after using PCA to reduce the dimension on raw data
- (2) Clean data - data after our data cleaning process, which set all values greater than 0 to 1 in our data
- (3) PCA Clean daata - data after using PCA to reduce the dimension on clean data after data cleaning process

### 4.1 Cross-Validation

When we build the model, it is normal to follow the principle of simplification since the simpler model we built, the better performance we will get. However, for some complicated problems, our model will also become more complex which might cause the overfitting problem. In order to solve this problem, we introduce the cross-validation technique. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it [16].

The purpose of cross-validation is to select the model with the optimal parameters. After the model is set up, tuning the parameters is a very time-consuming process. Through cross-validation, we can get the model with the optimal parameters much easier. Here are some steps about cross-validation procedure:

- (1) Prepare the candidate models,  $M_1, M_2, M_3, \dots$  (the model framework is consistent, only different on the parameters);
- (2) For each model, use cross-validation to return the accuracy and error rate information of the model, the result should be the average of cross-validation;

- (3) Select the best model by comparing the accuracy or error of the different models.

There are some types of cross-validation which are common to use: K-fold cross-validation and Leave-one-out cross-validation.

- K-fold:

This method is to divide the data set into  $k$  subsets. Each time, select one of the  $k$  subsets as the test set and the other  $k - 1$  subsets become a training set. Then the average accuracy or error across all  $k$  trials is computed [16]. In general, we choose 10 as the value of  $k$ .

- Leave-one-out (LOO):

This method is K-fold cross-validation taken to its logical extreme, with  $k = n$  ( $n > k$ ), the number of data points in the set [16]. That is, it randomly select  $n$  samples as a training set and the rest as a test set. Since the time complexity of this cross-validation is factorial, it is not an appropriate method for big data set.

In this project, we use K-fold cross validation technique to reduce over-fitting of our model and increase the accuracy in each model. We use the function `cross_val_score` from the `sklearn` package. It have several important parameters to set [10].

- (1) CV: int, cross-validation generator or an iterable, optional  
This parameter determines the cross-validation splitting strategy, which determined the number of fold we need to use. In our project, we use the default 3-fold cross-validation. Because 3-fold provide us the result in a reasonable time and accuracy.
- (2) Scoring: string, callable or None, optional, default: None  
The scoring parameter determines what to return after we call the function. We just this parameter to ‘accuracy’, which will return the accuracy between 0 to 1 for each model.

After we receive the result for each validation, we generate the mean of each result and use the result as the accuracy of the model.

```
from datetime import datetime
from sklearn.cross_validation import cross_val_score

def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label, cv=5,
                          scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds

    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)
```

**Figure 4: The core codes about cross-validation**

## 4.2 Decision Tree

### 4.2.1 Introduction

Decision tree builds classification or regression models in the form of a tree structure (either binary or non-binary) [20]. Each of its non-leaf nodes represents a test on the characteristic attributes, and each leaf node stores a category. The process of decision making using decision tree has the following steps [23]:

- (1) Start at the root node;
- (2) Test the corresponding characteristic attribute of the items that need to be classified;
- (3) Select the branch based on the value until the leaf node is reached;
- (4) The category of stored in the leaf node is the result.

The decision tree construction process rely on attribute selection metrics in order to choose the attribute which has the capability to divide tuples into different classes best. The key step in constructing a decision tree is split attributes which means to construct different branches according to the different partition of a certain characteristic attribute at a node. The goal of this step is to make each split subset as “pure” as possible. Split attributes are divided into three different situations:

- (1) Attributes are discrete values and do not require to generate a binary decision tree. This time, each partition of an attribute becomes a branch;
- (2) Attributes are discrete values and require to generate a binary decision tree. This time, a subset of attribute partitions is used for testing, broken down two branches according to “subordinate to this subset” and “not subordinate to this subset”;
- (3) Attributes are continuous values. This time, determine a value as a `split_point` and generate two branches according to  $> \text{split\_point}$  and  $\leq \text{split\_point}$ .

There are many attribute selection metric algorithms (e.g. ID3, C4.5, CART, etc.), generally using top-down recursive method with non-backtrack greedy strategy. In our experiment, we applied optimized version of the Classification And Regression Trees (CART) algorithm from scikit-learn library.

The CART algorithm uses a binary recursive segmentation technique [1]: the current sample set is divided into two sub-sample sets, so that each non-leaf node have two branches. Therefore, the decision tree generated by the CART algorithm is a concise binary tree with the root node represents a single input variable ( $x$ ) and a split point on that variable and the leaf nodes contain an output variable ( $y$ ) which has the capability to make a prediction [1].

The first key step of CART algorithm is creating the tree model, it examines each variable and all possible partitions of this variable to observe the best partitions. For discrete values such as  $U = \{x, y, z\}$ , there are three cases of partitions [9]:

$$\{\{x, y\}, \{z\}\}, \{\{x, z\}, \{y\}\}, \{\{y, z\}, \{x\}\}$$

except  $\emptyset$  and  $U$ ; for continuous values, it introduces the idea of “split point”. Suppose one attribute of a sample has  $n$  continuous values, it then has  $n - 1$  splitting points where each of them is the average of two consecutive values  $(a[i] + a[i + 1])/2$ . Partitions of each attribute are sorted by the amount of impurities that they can reduce. The reduction of impurities could use the most popular method of

impurity metric which is: Gini index. If we use  $k$  ( $k = 1, 2, 3, \dots, C$ ) to represent the class, where  $C$  is the dependent variable number of the category set. Thus, the Gini impurity of a Node  $A$  could be defined as [9]:

$$Gini(A) = 1 - \sum_{k=1}^C p_k^2$$

Where  $p_k$  denotes the probability of observation points which belong to class  $k$ . When  $Gini(A) = 0$ , all samples belong to the same class. When  $Gini(A)$  is the maximum, which is  $\frac{(C-1)C}{2}$ , all classes occur with the same probability in nodes.

The second key idea in the CART process is to prune the trees of the training set with independent validation data sets. Analyzing the recursive tree construction of classification and regression tree, it is easy to find that there exists a data over-fitting problem [1]. In the construction of decision tree, many branches reflect the abnormality in training data due to the noises or outliers inside. Using such decision tree to classify the data with unknown categories, the accuracy of classification is not high. So it is essential to detect and subtract these branches. Generally, tree pruning method uses statistical metrics, subtract the least reliable branches, which results in faster classification and improves the ability to separate correctly from the training data. The CART algorithm often adopts the post-pruning method, which is implemented by pruning the branches in a fully grown tree. By deleting the branch of the node to cut tree nodes, the bottom non-pruned node becomes a leaf.

The main codes of Figure 5 shows how we called CART algorithm in our experiment.

```
# Import Library
from sklearn import tree

def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")

    # Train the model using the training sets and check
    # score
    model_acc(data, label, dt_model)
```

**Figure 5: The core codes about CART algorithm (decision tree)**

#### 4.2.2 Advantage and Disadvantage

Decision Tree has advantages as follow [6]:

- (1) Decision trees are easy to understand and implement, and people have the ability to understand what the decision tree means by explaining it.
- (2) Data preparation is often simple or unnecessary for decision trees, and other techniques often require first generalizing data, such as removing redundant or blank attributes.
- (3) Feasible and effective results for large data sources in a relatively short period of time.

- (4) Not sensitive to missing values
- (5) Can handle irrelevant feature data
- (6) High efficiency. Decision tree only needs to build once. The maximum number of calculations for each prediction does not exceed the depth of the decision tree

Decision Tree also has disadvantages as follow [6]:

- (1) Hard to predict features with continues value
- (2) Need to do a lot of data reprocessing work for time-series data
- (3) When the category is too large, the error rate may increase.
- (4) It does not look good when dealing with data that has a strong correlation between each feature.

#### 4.2.3 Result

From table 1, we can find that the Decision Tree algorithm has a highest accuracy 0.8378 when we using Clean data. That's because the Clean data contains all 784 features in the data set. It has the minimum information loss among all three data set. Clean data also have the longest running time, which is 20 seconds.

PCA Clean Data have the second highest accuracy with the lowest running time. By using the PCA to reduce the dimension of the clean data, the running time reduced a lot. The accuracy only decreases by 0.01, which shows that the process of PCA did not lose a lot of information.

When we use decision tree algorithm, PCA data have the lowest accuracy. That's may because the raw data have may noise and redundant information. After we remove this information from our data pre-processing step, our accuracy increased.

### 4.3 Naive Bayes

#### 4.3.1 Introduction

Naive Bayes algorithm is a classification technique based on Baye's Theorem with an assumption of independence among predictors [18]. That is to say, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, we may guess a fruit is an orange if it is yellow, round and about 3 inches in diameter. Even if these features depend on each other, all properties independently contribute to the probability that this fruit is an orange, which explain the term 'Naive' [18].

The Baye's Theorem is particularly useful and not complicated. It solves many problems encountered in our life. The purpose of this theorem is that given a conditional probability of a certain condition, obtain the probability of exchanging two conditions. That is, to get  $P(B|A)$  while given  $P(A|B)$ .  $P(A|B)$  is the posterior probability which is also the conditional probability (likelihood) and  $P(A)$  or  $P(B)$  is called a prior probability. We use the following equation to express this theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The idea of Baye's Theorem is very simple and directly: For the given item which need to be classified, compute the probability of each category under this item. We consider this item belongs to the category with the largest value. The work process of Naive Bayes classification is as follows [25]:

	PCA Data	PCA Clean Data	Clean Data
Time	12	9	20
Accuracy	0.8012	0.8234	0.8378

Table 1: Result For Decision Tree

- (1) Let  $D$  be the set of training tuples associated with their class labels. Each tuple is represented by an  $n$ -dimensional attribute vector  $X = x_1, x_2, \dots, x_n$ ;
- (2) Suppose there are  $m$  classes  $C_1, C_2, \dots, C_m$ . For the given tuple  $X$ , the classification algorithm will predict that  $X$  belongs to the class with the highest posterior probability. That is, Naive Bayes classification predicts that  $X$  belongs to class  $C_i$  if and only if  $P(C_i|X) > P(C_j|X), 1 \leq j \leq m, j \neq i$ . Thus, the class  $C_1$  with the largest  $P(C_i|X)$  is called the maximum posterior probability according to the Baye's Theorem:  $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$ ;
- (3) Since  $P(X)$  is a constant for all classes, we only require the maximum of  $P(C_i|X)P(C_i)$ . If the prior probability of a class is unknown, then generally assume these classes are equiprobable (i.e.  $P(C_1) = P(C_2) = \dots = P(C_m)$ ) and maximize  $P(C_i|X)$  based on this assumption. Otherwise, maximize  $P(C_i|X)P(C_i)$ ;
- (4) Given a data set with multiple attributes, the computational cost of  $P(C_i|X)$  is very large. In order to reduce this cost, we could make the naive assumption about conditional independent of the class. For the label of a given tuple class, assuming the attribute values are conditionally independent. Therefore, we have

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

To examine whether the attribute is classified or continuous value, we need to consider the following two cases:

- (a) If  $A_k$  is a classified attribute, then  $P(x_k|C_i)$  is the number of tuples of class  $C_i$  whose value is  $x_k$  for attribute  $A_k$  in  $D$  divided by the number of tuples of class  $C_i$  in  $D$  ( $|C_i, D|$ );
- (b) If  $A_k$  is a continuous value attribute, then assume the attribute obeys a Gaussian distribute with the mean  $\eta$  and standard deviation  $\sigma$ , as defined by:

$$g(x, \eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

Thus,  $P(x_k|C_i) = g(x_k, \eta_{C_i}, \sigma_{C_i})$ .

- (5) To predict the label of class  $X$ , calculate  $P(C_i|X)P(C_i)$  for each class  $C_i$ .

The whole Naive Bayes classification could be divided into three stages:

- (1) Preparation stage. The task of this stage is to make the necessary preparation for the Naive Bayes classification. The main work is to determine the characteristic attributes according to the specific situations and make the appropriate partition for each characteristic attribute, and then

manually classified some of the items to constitute a training sample set. The input of this stage is all data that need to be classified and the output is the characteristic attribute and the training sample.

- (2) Classifier training stage, the task of this stage is to generate a classifier. The main work is to compute the occurrence frequency of each class in training sample and the conditional probability of all partitions in each category, and then record the results. The input is characteristic attributes and a training sample, the output is a classifier. This stage could be completed automatically by a program.
- (3) Application stage. The task of this stage is to classify items using classifier. The input is classifier and items, and the output is the mapping between items and categories. This stage could also be completed by a program.

The main codes of Figure 6 shows how we called Naive Bayes algorithm in our experiment.

```
# Import Library
from sklearn.naive_bayes import GaussianNB

def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")

    # Train the model using the training sets and check
    # score
    model_acc(data, label, nb_model)
```

Figure 6: The core codes about Naive Bayes

#### 4.3.2 Advantage and Disadvantage

Naive Bayes has advantages as follow [13]:

- (1) Naive Bayesian model originated in classical mathematical theory, which is stable.
- (2) Have a good performance on small-scale data,
- (3) Can handle multi-category tasks.
- (4) For incremental training, especially when the amount of data exceeds memory, we can use batch training to save training time.

Naive Bayes also has disadvantages as follow [13]:

- (1) In theory, the naive Bayes model has the smallest error rate compared to other classification methods. However, this is not always the case. This is because the naive Bayesian model assumes that the features are independent of each other. This assumption often does not hold in practice.

- When the number of attributes is large or the correlation between attributes is large, the error rate will be huge.
- (2) Need to know the prior probability, and the probability of prior probability depends on the assumption. There are many kinds of hypothetical models, so the prediction results will be poor at some time due to the choice of hypothetical model.
  - (3) Because we determine the posterior probability by priority and data to determine the classification, there is a certain error rate in the classification decision.
  - (4) Sensitive to the type of raw data.

#### 4.3.3 Result

From table 2, we can find that Clean Data have a really low accuracy with the highest time spent. That's because the raw data set did not match the assumption of Naive Bayes. The features are not conditionally independent of each other. The pixels are continues. For example, if pixel1 and pixel3 are both greater than 0, pixel2 will have a more probability to have a value greater than 0.

After we use the dimension reduction technique to reduce the dimension of the data, each component of the data becomes a linear combination of the original data. The new data fits the assumption of Naive Bayes more. Therefore, the PCA Data and PCA Clean Data have a much better performance than Clean Data. They also have the lowest running time compare to any other algorithms.

The PCA Clean Data have the highest accuracy of 0.8710 which higher than the PCA Data. That's may because of the noise and redundant in the original data.

## 4.4 Logistic Regression

### 4.4.1 Introduction

Logistic regression is a static regression model with a category of the dependent variable. It uses a binary logistic model to estimate binary response probability on predictor variables. In this case, we can know which specific factor makes influence in the presence of risk increasing odds when getting outcomes. We use logistic regression to find the best fitting model to conclude the relationship between variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of the presence of the characteristic of interest [22]:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

$p$  is the probability of the presence of the characteristic of interest and odds is logical transformation.

$$\text{odds} = \frac{p}{1-p} = \frac{p(\text{presence of characteristic})}{p(\text{absence of characteristic})}$$

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

There are four ways to input independent variables into the model:

- (1) Enter: enter all variables at the same time
- (2) Forward: enter essential variables one by one
- (3) Backward: enter all variables first and delete non-essential variables one by one

- (4) Stepwise: enter essential variables one by one and check the importance of each variable, delete non-essential ones.

It still has other options:

- (1) Remove variable. Variables would be removed from the model if its significant level is greater than P-value.
- (2) Classification table cutoff value: a value between 0 and 1 which will be used as a cutoff value for a classification table. The classification table is a method to evaluate the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at the selected cut-off value) are cross-classified [22].
- (3) Categorical: Identify variables in the category.

The main codes of Figure 7 shows how we called Logistic Regression algorithm in our experiment.

```
# Import Library
from sklearn.linear_model import LogisticRegression

def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")

    # Train the model using the training sets and check
    # score
    model_acc(data, label, lr_model)
```

Figure 7: The core codes about Logistic Regression

### 4.4.2 Advantage and Disadvantage

Logistic Regression has advantages as follow [14]:

- (1) Very simple to implement and use, widely used in industrial issues
- (2) The amount of computation is very small when classified. Therefore the running time is low and the requirement for the storage space is also low.
- (3) The sigmoid score for each sample is easy to observe. The threshold can be easily determined by user.
- (4) For logistic regression, multicollinearity is not a problem, it can be solved in conjunction with L2 regularization;

Logistic Regression also has disadvantages as follow [14]:

- (1) When the feature space is large, the performance of logistic regression is not very good.
- (2) May have the under-fitting problem, the general accuracy is not high.
- (3) Can only deal with the binary classification problem (based on this, softmax can be used for multi-classification), and must be linearly separable.
- (4) For non-linear features, normalization is required.

### 4.4.3 Result

The result of logistic regression is pretty impressive. This is a 10-categorical classification problem, and logistic regression did a good job on this task.

	PCA Data	PCA Clean Data	Clean Data
Time	0	0	20
Accuracy	0.8651	0.8710	0.5397

Table 2: Result For Navie Bayes

	PCA Data	PCA Clean Data	Clean Data
Time	27	21	218
Accuracy	0.8891	0.8862	0.9064

Table 3: Result For Logistic Regression

When we get this result, we are thinking if we having an over-fitting result. Therefore, we add a regularization parameter to penalize the features. We use l2 regularization as our parameter when we create our logistic classifier. We also use cross-validation skill to increase our sample size. The results show that the accuracy is still around 90%. Therefore, we are not having an over-fitting problem.

The running time of logistic regression is relatively high. For Clean Data, it received the accuracy of 0.9064 with 218 seconds. PCA Data and PCA Clean Data have a lower accuracy with a much lower time spend. Also, we noticed that the PCA Data accuracy is a little bit higher than the PCA Clean Data. That's may because the clean data make some of the information loss in the raw data.

## 4.5 Random Forest

### 4.5.1 Introduction

Random forest uses a random way to build a forest within many decision trees. There is no correlation between each tree in a random forest [26]. After getting the forest, when a new input sample comes in, each decision tree required to make a judgment separately in order to see which class the sample belongs to (for the classification algorithm), and predict the sample for the category which has most selected.

Random forest is mainly used for regression and classification. It is somewhat similar to the bagging which utilizes decision trees as a basic classifier. Bagging could generate a decision tree after replay a sample in each bootstrap and do not make more intervention while generating these trees. Random forest is also sampling with bootstrap, but the difference is that when constructing each tree, every node variable is generated only in a small number of randomly selected variables. Therefore, not only the samples are random, but also the generation of each node's features. Since the combination classifier is more effective than the single classifier, random forest could classify the data and give the importance evaluation of each variable.

The basic principle of random forest is to get a new training sample set by selecting  $k$  samples from the original training sample set  $N$ , and then make up a random forest according to  $k$  classification trees. The classification result of the new data depends on the score of the tree votes [17]. In essence, it is an improvement on the decision tree algorithm: it combines multiple decision trees, each tree established depends on an independently sample and has the same distribution. The classification error relies on each the classification ability of a tree and the correlation between them.

Feature selection uses a random method to split each node, and then compare the error generated in different situations. The inherent estimation error, classification ability and relevance determine the number of features [17].

Since there are many decision trees in the forest, once a new input sample comes in, each decision tree make a decision to check what the class the sample belongs to, and which one is chosen most to the prediction. There are two selection metrics for decision trees to split attributes [23]:

- (1) Information gain
    - (a)  $I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^m p_i \log_2(p_i)$ , where  $S$  is the data set,  $m$  is the number of categories,  $p_i \approx \frac{|S_i|}{|S|}$  is the probability for any sample belongs to  $C_i$ ,  $C_i$  is a class label and  $s_i$  is the number of samples on  $C_i$ ;
    - (b) The smaller  $I(s_1, s_2, \dots, s_m)$ , the more ordered of the sample and the better the classification effect;
    - (c) Entropy of the subsets partitioned by attribute  $A$ :  $A$  has  $V$  different values,  $S$  is partitioned by  $A$  into  $V$  subsets  $s_1, s_2, \dots, s_V$ , where  $s_{ij}$  is the number of samples of  $C_i$  in subset  $s_j$ . Then, we have
  - (d)  $E(A) = \sum_{j=1}^V \frac{(s_{1j} + \dots + s_{mj})}{s * I(s_{1j}, \dots, s_{mj})}$
  - (e)  $G = I(s_1, s_2, \dots, s_m)E(A)$ ;
  - (f) Select the attribute with the maximum information gain as the split attribute.
- (2) Gini index
    - (a) Set  $S$  contains  $N$  categories of records, then its Gini index is the frequency of the occurrence of  $p_j$ ;
    - (b) If set  $S$  is partitioned into  $m$  parts  $s_1, s_2, \dots, s_m$ , this segmentation is the Gini split;
    - (c) Select the attribute with the smallest Gini split as a split attribute.

In order to implement random forest, we should follow these steps:

- (1) The input original training set is  $N$ , use bootstrap to extract  $k$  samples randomly and build  $k$  decision trees;
- (2) Suppose there are  $m_A$  variables, then randomly extract  $m_T$  variables from each node of each tree to find one of the variables with the highest classification ability in  $m_T$  variables. The threshold of the variable classification is determined by checking each classification point;

- (3) Maximize the growth of each tree without any pruning;
- (4) Constitute the random forest with these decision trees. Use random forest to determine and classify the new data, and the results are based on votes amount of the tree classifier.

The main codes of Figure 8 shows how we called Random Forest algorithm in our experiment.

```
# Import Library
from sklearn.ensemble import RandomForestClassifier

def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")

    # Train the model using the training sets and check
    # score
    model_acc(data, label, rf_model)
```

**Figure 8: The core codes about Random Forest**

#### 4.5.2 Advantage and Disadvantage

Random Forest has advantages as follow [2]:

- (1) It can handle very high-dimensional data, and do not have to do feature selection, feature subset is randomly selected
- (2) It can provide which feature is more important after training.
- (3) When creating a random forest, the use of generalization error is an unbiased estimation, which shows that this model has a high generalization ability.
- (4) Easy to make a parallel method, training tree and tree are independent of each other.
- (5) In the training process, the algorithm is able to detect the interaction between the features.
- (6) For unbalanced data sets, it can balance the model automatically.
- (7) If a large part of the features is lost, the model can still maintain the accuracy.

Random Forest also has disadvantages as follow [2]:

- (1) There may be many similar decision trees that mask the real results.
- (2) Small data or low dimensional data may not produce the best classification.
- (3) Much slower than single decision tree algorithm.
- (4) Random forests can be over-fitting on some noisy classifications or regression problems
- (5) For feature with different value range, the more value-separated features will have a greater impact on random forests

#### 4.5.3 Result

From table 4, we can find that Clean Data performed perfectly in this case. It takes the shortest time and reached a 0.9647 accuracy.

The result shows an interesting phenomenon: Clean Data cost less time than PCA Data and PCA Clean Data. In order to explain this phenomenon, we have to check what parameter we choose when we build our random forest classifier. From sklearn API document, we can find that the first default parameter is the number of trees in the forest. For all the data, we set the number of trees to the default number, which is 10. However, in Clean Data, many features are correlated to each other, which means that there may many similar decision trees. For PCA Data and PCA Clean Data, most of the features are independent of each other. Therefore, the running time for Clean Data is higher than PCA Data and PCA clean Data.

Also, we know that when there are similar decision trees in the random forest, the real results may be masked. Therefore, although the Clean Data have a really high accuracy, it may still not as good as the PCA Data and PCA Clean Data result. When we running the classifier on an untested data set, the classifier made by PCA Clean Data may have the best performance among the three.

## 4.6 Support Vector Machine

### 4.6.1 Introduction

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [27]. It is mostly used in classification. People can plot each data as a point in an n-dimensional space and give each feature a value. Finding the hyperplane which can differentiate two classes very well can complete classification. As for hyperplane, we must know the notation used to define a hyperplane [15]:

$$f(x) = \beta_0 + \beta^T x$$

$\beta$  is weight and  $\beta_0$  is bias. The optimal hyperplane can be represented in an infinite number of different ways by scaling of  $\beta$  and  $\beta_0$ . The one we choose is [15]:

$$|\beta_0 + \beta^T x| = 1$$

$x$  is the training sample who is the most closest to hyperplane. It is known as canonical hyperplane. Distance between point and hyperplane is [15]:

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}$$

$$\text{distance}_{\text{support vector}} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|}$$

$$M = 2 * \text{distance}_{\text{support vector}} = \frac{2}{\|\beta\|}$$

$$\min L(\beta) = \frac{1}{2} \|\beta\|^2 \text{ subject to } y_i(\beta^T + \beta_0) \geq 1, \forall i$$

In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below Python code [19] in Figure 9:

The e1071 package in R is used to create Support Vector Machines with ease. It has helper functions as well as code for the Naive Bayes Classifier. The creation of a support vector machine in R and Python

	PCA Data	PCA Clean Data	Clean Data
Time	126	107	56
Accuracy	0.9483	0.9497	0.9647

**Table 4: Result For Random Forest**

```

# Import Library
from sklearn import svm

# Assumed you have, X (predictor) and Y (target) for
# training data set and x_test(predictor) of test_data
# set

# Create SVM classification object
model = svm.SVC(kernel='rbf', C=10)

# there are various option associated with it, like changing
# kernel, gamma and C value
# Train the model using the training sets and check score
model.fit(X, y)
model.score(X, y)

# Predict Output
predicted= model.predict(x_test)

```

**Figure 9: The core codes about SVM in Python**

follow similar approaches, let's take a look now at the following R code [19] in Figure 10:

```

# Import Library
require(e1071) #Contains the SVM

Train <- read.csv(file.choose())
Test <- read.csv(file.choose())

# there are various options associated with SVM training;
# like changing kernel, gamma and C value.

# create model
model <-
  svm(Target~Predictor1+Predictor2+Predictor3,data=Train,
  kernel='linear',gamma=0.2, cost=100)

# Predict Output
preds <- predict(model,Test)
table(preds)

```

**Figure 10: The core codes about SVM in R**

#### 4.6.2 Advantage and Disadvantage

Support vector machine has advantages as follow:

- (1) More efficient in high dimensional space.
- (2) Effective when the number of samples is smaller than the number of dimensions.
- (3) Can memorize efficiently by using a subset of training sample in decision function.
- (4) Flexible by changing Kernel functions for different customers.

And it also has disadvantages as follow:

- (1) It would over-fitting in choosing Kernel functions when the number of samples is much smaller than the number of features.
- (2) Must pay more attention to regularization term.
- (3) It can only get probability by an expensive five-fold cross-validation instead of calculating directly.

#### 4.6.3 Result

By using SVM to build our classifier, we received a really great accuracy score. The PCA Data received a 0.9814 accuracy of 127 seconds. The running complexity of SVM is  $O(N^3 + LN^2 + d * L * N)$ , which  $N$  is the number of support vector choose,  $L$  is the number of samples and  $d$  is the number of features of the data set. Therefore, SVM algorithm will run really slow on the large data set. Therefore, when we use the Clean Data, which include more than 42000 samples and 784 features, it takes 1029 seconds to finish the job. We also try to use SVM direct on our raw data. It takes forever to get a result.

SVM can get much better results than other algorithms in the small sample training set. SVM has become one of the most commonly used and effective classifiers. By using the concept of margin, a structured description of the data distribution is obtained, thereby reducing the need for data size and data distribution.

SVM model has three very important parameters kernel,  $C$  and  $\gamma$  [11].

- (1) Kernel: string, optional. This parameter specifies the kernel type to be used in the algorithm. There are many different kernels that can be used in SVM. For example, linear, polynomial, sigmoid, Radial basis function (RBF) and pre-computed. In this project, choose to use RBF. Because:
  - (a) The RBF kernel function can map a sample to a higher dimensional space, and the linear kernel function is a special case of RBF. That is to say, if RBF is considered, then it is unnecessary to consider the linear kernel function.
  - (b) Compared with polynomial kernel function, RBF needs to determine fewer parameters, the number of kernel function parameters directly affect the complexity of the function. In addition, when the order of the polynomial is relatively high, the elemental values of the

	PCA Data	PCA Clean Data	Clean Data
Time	127	90	1029
Accuracy	0.9814	0.9785	0.9575

**Table 5: Result For Support Vector Machine**

kernel matrix will tend to positive infinity or negative infinity, while the RBF will reduce the numerical calculation difficulties.

- (c) RBF and sigmoid have similar performance for some parameters.
- (2) C is the penalty coefficient, which shows the tolerance of the bias. If your C is small, it will give you a great distance, but as a trade-off, we have to ignore some misclassified samples; on the other hand, if you have a large C, you will try to correctly classify all the samples, but the price is the margin space will be small. In our example, we choose c equals to 10, which is a relatively large c value, which brings us a more accurate classifier.
- (3) Gamma defines how much influence a single training example has. It determines the distribution of the data after mapping to a new feature space. The larger the gamma is, the less the support vector it will be. The smaller the gamma value is, the more the support vector it will be. The number of support vectors affects the speed of training and prediction. Also, if we set gamma large, it will have the over-fitting problem. Therefore, in this project, we decided to use the default gamma value, which is

$$\gamma = \frac{1}{\text{number of features}}$$

In this task, SVM have a really great performance, the running time is also acceptable.

## 5 CONCLUSION

From table 6, we can easily find that when we use SVM classifier on PCA Data, we will receive the highest accuracy among all 5 different algorithms. The highest accuracy we reached for this project is 0.9813, which shows that our classifier predicts 98.13% of the sample correct by using our SVM classifier. The time of training the model takes 127 seconds. The time spent is acceptable. The accuracy of SVM on PCA Clean Data has the second highest accuracy, which is 0.9785. The difference between first and second highest accuracy is about 0.0028, which is really small. However, the time spent saved 41.1%. Therefore, SVM on PCA Clean Data is also a reasonable choice for the Digit Recognition task.

Random Forest can be explained as a combination of many decision trees. Decision tree can be explained as a special case of Random Forest, which set the number of trees in the Random Forest to 1. Therefore, Random Forest has a much better performance than decision tree in all three data set. As a trade-off, the time spent for Random Forest is much higher than Decision Tree.

Compare to other four Classifiers, Naive Bayes has the fastest training speed. For PCA Data and PCA Clean Data, Naive Bayes Classifier takes less than one second to train the classifier. And for

Clean data, which contains all 784 features, it takes only 6 seconds to train the classifier. The reason why Naive Bayes is fast is that:

- (1) The algorithm does not need to iterate to get the result. The running time is approximately linear.
- (2) It makes an assumption of independence between its features, so that parameter estimates can be calculated independently and thus possibly very quickly.
- (3) The prior probability values do not change. Therefore, the prior probability can be calculated and store in memory in the first place.

However, we have to be very careful about the assumption made by Naive Bayes, or we will get a very low accuracy.

Logistic Regression received an average performance among the 5 algorithms. It achieves a 0.8891 accuracy in 27 seconds on PCA data. However, when we using logistic regression, we have to pay a lot of attention to over-fitting problem. We should use regularization and cross-validation to reduce the probability of over-fitting problems.

To conclude, we decide to use SVM classifier for Digit Recognition Task. We should definitely use feature extraction on the data because of the running time and over-fitting problem. The Binary Data cleaning method is optional. If we want to have higher accuracy, we should not use Binary Data cleaning. As a trade-off, if we want to have faster training speed, we should use Binary Data cleaning.

## 6 LIMITATIONS

Our analysis is far from perfect. There are several points that we want to point our as discussion and also opportunities for future improvement.

- (1) We can try several more classification algorithms. For example,  $K^{\text{th}}$  Nearest Neighbour (KNN) and Neural Network. We can use some more complex algorithms too, such as Convolution Neural Network (CNN).
- (2) We can focus more on tune parameter. For example, we can use the Grid Search on SVM to get a better parameter combination.
- (3) We can choose a different Data Cleaning Method. For example, we can set a threshold on data. Any value greater than 50 will be set to 1.
- (4) We can choose a different Feature Extraction or Feature Selection method. For example, LDA. Unlike PCA, LDA is an unsupervised dimension reduction method.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski and all TAs for providing the resource, tutorials and other related materials to write this paper.

	Decision Tree	Naive Bayes	Logistic Regression	Random Forest	Support Vector Machine
PCA Time	12	<b>0</b>	27	126	127
PCA Accuracy	0.8012	0.8651	0.8891	0.9483	<b>0.9813</b>
PCA Clean Time	9	<b>0</b>	21	107	90
PCA Clean Accuracy	0.8234	0.8710	0.8862	0.9497	0.9785
Clean Time	20	6	218	56	1029
Clean Accuracy	0.8378	0.5397	0.9064	0.9647	0.9575

Table 6: Result For Different Algorithm with Different Data Cleaning & Feature Extraction method

## REFERENCES

- [1] Jason Brownlee. 2016. Classification And Regression Trees for Machine Learning. (April 2016). <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>
- [2] Daniel S. Chapman, Aletta Bonn, William E. Kunin, and Stephen J. Cornell. 2009. Random Forest characterization of upland vegetation and management burning from aerial imagery. *Journal of Biogeography* 37, 1 (2009), 37–46. <https://doi.org/10.1111/j.1365-2699.2009.02186.x>
- [3] Madasu Hanmandlu, KR Murali Mohan, Sourav Chakraborty, Sumeer Goyal, and D Roy Choudhury. 2003. Unconstrained handwritten character recognition based on fuzzy logic. *Pattern Recognition* 36, 3 (2003), 603–623.
- [4] J. Hobcraft and Bernard Benjamin. 1970. The Population Census. *Population Studies* 24, 3 (1970), 460. <https://doi.org/10.2307/2173052>
- [5] Kazuhiro Hotta. 2008. Non-linear feature extraction by linear PCA using local kernel. *2008 19th International Conference on Pattern Recognition* (2008). <https://doi.org/10.1109/icpr.2008.4761721>
- [6] Hemant Ishwaran and J. Sunil Rao. 2009. Decision Tree: Introduction. *Encyclopedia of Medical Decision Making* (2009). <https://doi.org/10.4135/9781412971980.n97>
- [7] Kaggle. 2015. Data Discription. (2015). <https://www.kaggle.com/c/digit-recognizer/data>
- [8] C. Kamath and R. Musick. 1998. Scalable pattern recognition for large-scale scientific data mining. (1998). <https://doi.org/10.2172/310913>
- [9] Scikit Learn. 2007. Decision Trees. (2007). <http://scikit-learn.org/stable/modules/tree.html>
- [10] Scikit Learn. 2007. sklearn model selection cross val score. (2007). [http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.cross\\_val\\_score.html](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html)
- [11] Scikit Learn. 2007. sklearn svm SVC. (2007). <http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
- [12] Sheng-Jie Liang, Zhi-Hua Zhang, and Li-Lin Cui. 2010. Feature extraction method Based PCA and KICA. *2010 Second International Conference on Computational Intelligence and Natural Computing* (2010). <https://doi.org/10.1109/cinc.2010.5643821>
- [13] J. Luengo and Rafael Rumi. 2015. Naive Bayes Classifier with Mixtures of Polynomials. *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (2015). <https://doi.org/10.5220/0005166000140024>
- [14] Scott Menard. 2010. Introduction: Linear Regression and Logistic Regression. *Logistic Regression: From Introductory to Advanced Concepts and Applications* (2010), 1–18. <https://doi.org/10.4135/9781483348964.n1>
- [15] OpenCV. 2017. Introduction to Support Vector Machines. (December 2017). [https://docs.opencv.org/2.4/doc/tutorials/ml/introduction\\_to\\_svm/introduction\\_to\\_svm.html](https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html)
- [16] OpenML. 2013. 10-fold Crossvalidation. (2013). <https://www.openml.org/a/estimation-procedures/1>
- [17] Savan Patel. 2017. Chapter 5: Random Forest Classifier. (May 2017). <https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1>
- [18] Sunil Ray. 2015. 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R). (September 2015). <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
- [19] Sunil Ray. 2015. Understanding Support Vector Machine algorithm from examples (along with code). (October 2015). <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
- [20] Dr. Saed Sayad. 2010. Decision Tree - Classification. (2010). <http://www.saedsayad.com/decision.tree.htm>
- [21] Faisal Tehseen Shah and Kamran Yousaf. 2007. Handwritten Digit Recognition Using Image Processing and Neural Networks. *Proceedings of the World Congress on Engineering* (July 2007).
- [22] MedCalc Software. 2017. Logistic regression. (February 2017). [https://www.medcalc.org/manual/logistic\\_regression.php](https://www.medcalc.org/manual/logistic_regression.php)
- [23] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Addison Wesley.

- [24] Tong-qing WANG, Yan JU, and Li RENG. 2003. Handwritten Digit Recognition Based on Neural Networks and Multi-structure Information Fusion [J]. *Micro Systems* 12 (2003), 059.
- [25] Wikipedia. 2017. Naive Bayes classifier. (December 2017). [https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)
- [26] Wikipedia. 2017. Random forest. (November 2017). [https://en.wikipedia.org/wiki/Random\\_forest](https://en.wikipedia.org/wiki/Random_forest)
- [27] Wikipedia. 2017. Support vector machine. (December 2017). [https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine)
- [28] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar. 2005. Enhancing Data Analysis with Noise Removal. (2005). <https://doi.org/10.21236/ada439494>

## A CODE ATTACHMENT

```

##Author: Yuchen Liu HID213, Wen Xuanhan HID209, Junjie Lu
##ID: 214
##Data: 2017.12.01
##Reference:
http://blog.csdn.net/tinkle181129/article/details/55261251

from datetime import datetime
import matplotlib.pyplot as plt
import pandas as pd
from numpy import *
from sklearn import svm
from sklearn import tree
from sklearn.cross_validation import cross_val_score
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB

# 1. read data from csv
def read_data():
    data_set = pd.read_csv("train.csv")
    data = data_set.values[0:, 1:]
    label = data_set.values[0:, 0]
    print("Data load completed.")
    return data, label

# plot 70 samples
def show_pic(data):
    print(shape(data))
    plt.figure(figsize=(7, 7))
    for digit_num in range(0, 70):
        plt.subplot(7, 10, digit_num + 1)
```

```

grid_data = data[digit_num].reshape(28, 28)
plt.imshow(grid_data, interpolation="none",
           cmap="afmhot")
plt.xticks([])
plt.yticks([])
plt.tight_layout()
plt.savefig("data_samples.png")

# 2. Data Cleaning
# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0
    print("Data clean completed.")
    return new_data

# 3. Feature Selection by PCA
def feature_selection(data):
    # First, use explained_variance to get recommended
    # number of component
    pca = PCA()
    # pca_parameter = pca.fit(data)
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            # print(n)
            break

    # Then, PCA the model by the number of components
    # pca = PCA(n_components=n, whiten=True)
    pca = PCA(n_components=n, whiten=True)
    print("Feature selection completed.")
    return pca.fit_transform(data)

# 4. Model Selection
def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label,
                          scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds
    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)

def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")
    model_acc(data, label, dt_model)

def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")
    model_acc(data, label, nb_model)

def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")
    model_acc(data, label, lr_model)

def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")
    model_acc(data, label, rf_model)

def svm_classifier(data, label, flag):
    svm_model = svm.SVC(kernel="rbf", C=10)
    svm_model.fit(data, label)
    # svc_clf = NuSVC(nu=0.1, kernel='rbf', verbose=True)
    print("Test " + flag + " using SVM: ")
    model_acc(data, label, svm_model)

def main():
    data, label = read_data()
    # show_pic(data)
    clean_data = data_clean(data)

    test_type = 3
    for i in range(1, 3):
        print("In %d test" % i)

        if test_type == 0:
            input_data = data
            str = "raw data"
        elif test_type == 1:

```

```
    input_data = clean_data
    str = "clean data"
elif test_type == 2:
    input_data = feature_selection(data)
    str = "pca data"
elif test_type == 3:
    input_data = feature_selection(clean_data)
    str = "pca clean data"

dt_classifier(input_data, label, str)
nb_classifier(input_data, label, str)
lr_classifier(input_data, label, str)
rf_classifier(input_data, label, str)
svm_classifier(input_data, label, str)

main()
```

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--entry type for "cart" isn't style-file defined  
--line 5 of file report.bib  
Warning--entry type for "sklearn.cv" isn't style-file defined  
--line 57 of file report.bib  
Warning--entry type for "sklearn.dt" isn't style-file defined  
--line 68 of file report.bib  
Warning--entry type for "svm.form" isn't style-file defined  
--line 117 of file report.bib  
Warning--entry type for "10f.cv" isn't style-file defined  
--line 129 of file report.bib  
Warning--entry type for "sp.rfc" isn't style-file defined  
--line 139 of file report.bib  
Warning--entry type for "nb.steps" isn't style-file defined  
--line 150 of file report.bib  
Warning--entry type for "svm.code" isn't style-file defined  
--line 162 of file report.bib  
Warning--entry type for "ss.dt" isn't style-file defined  
--line 174 of file report.bib  
Warning--entry type for "lr.form" isn't style-file defined  
--line 185 of file report.bib  
Warning--entry type for "wiki.nb" isn't style-file defined  
--line 208 of file report.bib  
Warning--entry type for "wiki.rf" isn't style-file defined  
--line 219 of file report.bib  
Warning--entry type for "wiki.svm" isn't style-file defined  
--line 231 of file report.bib  
Warning--no number and no volume in PCA  
Warning--page numbers missing in both pages and numpages fields in PCA  
Warning--no number and no volume in DT  
Warning--page numbers missing in both pages and numpages fields in DT  
Warning--no journal in app2  
Warning--no number and no volume in app2  
Warning--page numbers missing in both pages and numpages fields in app2  
Warning--no number and no volume in feature\_extra  
Warning--page numbers missing in both pages and numpages fields in feature\_extra  
Warning--no number and no volume in NB  
Warning--page numbers missing in both pages and numpages fields in NB  
Warning--no number and no volume in LR

```
Warning--no number and no volume in hdwc
Warning--page numbers missing in both pages and numpages fields in hdwc
Warning--empty address in intro.dm
Warning--no journal in data_clean
Warning--no number and no volume in data_clean
Warning--page numbers missing in both pages and numpages fields in data_clean
(There were 31 warnings)
```

```
bibtext _ label error
```

```
=====
```

```
report.bib:243:@Article{data_clean,
report.bib:89:@Article{feature_extra,
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
report.bib:252:@article{app1, title={The Population Census.},
report.bib:263:@article{app2, title={Scalable pattern recognition for large-scale scient
report.bib:268:@article{99, title={Unconstrained handwritten character recognition based
report.bib:278:@article{second, title={Handwritten Digit Recognition Based on Neural Net
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-12-16 09.32.30] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 1.6s.
./README.yml
24:81      error    line too long (82 > 80 characters)  (line-length)
47:79      error    trailing spaces  (trailing-spaces)
48:81      error    line too long (82 > 80 characters)  (line-length)
48:82      error    trailing spaces  (trailing-spaces)
49:79      error    trailing spaces  (trailing-spaces)
50:81      error    line too long (81 > 80 characters)  (line-length)
50:81      error    trailing spaces  (trailing-spaces)
51:81      error    line too long (89 > 80 characters)  (line-length)
51:89      error    trailing spaces  (trailing-spaces)
52:81      error    line too long (81 > 80 characters)  (line-length)
52:81      error    trailing spaces  (trailing-spaces)
```

```
53:81    error    line too long (81 > 80 characters)  (line-length)
53:81    error    trailing spaces  (trailing-spaces)
54:81    error    line too long (86 > 80 characters)  (line-length)
54:86    error    trailing spaces  (trailing-spaces)
```

---

Compliance Report

---

```
name: Han, Wenxuan
hid: 209
paper1: Oct 29 2017 100%
paper2: 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
14
wc 209 project 14 9520 report.tex
wc 209 project 14 9572 report.pdf
wc 209 project 14 853 report.bib
```

```
find "
```

---

```
169: print("Data clean completed.")

268: acc = cross_val_score(model, data, label, cv=5,
                           scoring="accuracy").mean()

272: print("Time use: ", time_use)

273: print("Accuracy by cross validation: ", acc)

325: print("Test " + data_type + " using DT: ")

423: print("Test " + data_type + " using NB: ")

509: print("Test " + data_type + " using LR: ")
```

```
603: print("Test " + flag + " using RF: ")  
850: data_set = pd.read_csv("train.csv")  
853: print("Data load completed.")  
864: plt.imshow(grid_data, interpolation="none", cmap="afmhot")  
868: plt.savefig("data_samples.png")  
883: print("Data clean completed.")  
910: print("Feature selection completed.")  
917: acc = cross_val_score(model, data, label,  
    scoring="accuracy").mean()  
920: print("Time use: ", time_use)  
921: print("Accuracy by cross validation: ", acc)  
927: print("Test " + data_type + " using DT: ")  
934: print("Test " + data_type + " using NB: ")  
941: print("Test " + data_type + " using LR: ")  
948: print("Test " + flag + " using RF: ")  
953: svm_model = svm.SVC(kernel="rbf", C=10)  
956: print("Test " + flag + " using SVM: ")  
967: print("In %d test" % i)  
971: str = "raw data"  
974: str = "clean data"  
977: str = "pca data"  
980: str = "pca clean data"  
passed: False
```

```
find footnote
-----
passed: True

find input{format/i523}
-----
6: \%input{format/i523}

passed: True

find input{format/final}
-----
7: \input{format/final}

passed: True

floats
-----
121: \begin{figure} [!ht]
123: \includegraphics[width=\columnwidth]{images/data_samples}
150: In our experiment, we selected method (2) to clean the raw data.
     The main codes in Figure \ref{fig:data clean} shows this
     operation.
152: \begin{figure} [htb]
172: \caption{The core codes about data clean}\label{fig:data clean}
189: \begin{figure} [htb]
214: \caption{The core codes about PCA processing}\label{fig:pca
     process}
260: \begin{figure} [htb]
275: \caption{The core codes about cross-validation}\label{fig:cv}
314: The main codes of Figure \ref{fig:dt} shows how we called CART
     algorithm in our experiment.
316: \begin{figure} [htb]
330: \caption{The core codes about CART algorithm (decision
     tree)}\label{fig:dt}
412: The main codes of Figure \ref{fig:nb} shows how we called Naive
     Bayes algorithm in our experiment.
414: \begin{figure} [htb]
428: \caption{The core codes about Naive Bayes}\label{fig:nb}
498: The main codes of Figure \ref{fig:lr} shows how we called
     Logistic Regression algorithm in our experiment.
500: \begin{figure} [htb]
```

```

514: \caption{The core codes about Logistic Regression}\label{fig:lr}
592: The main codes of Figure \ref{fig:rf} shows how we called Random
Forest algorithm in our experiment.
594: \begin{figure}[htb]
608: \caption{The core codes about Random Forest}\label{fig:rf}
673: In Python, scikit-learn is a widely used library for implementing
machine learning algorithms, SVM is also available in the scikit-
learn library and follows the same structure (Import library,
object creation, fitting model and prediction). Let's look at the
below Python code \cite{svm.code} in Figure \ref{fig:svm_p}:
675: \begin{figure}[htb]
694: \caption{The core codes about SVM in Python}\label{fig:svm_p}
697: The e1071 package in R is used to create Support Vector Machines
with ease. It has helper functions as well as code for the Naive
Bayes Classifier. The creation of a support vector machine in R
and Python follow similar approaches, let's take a look now at
the following R code \cite{svm.code} in Figure \ref{fig:svm_r}:
699: \begin{figure}[htb]
718: \caption{The core codes about SVM in R}\label{fig:svm_r}

```

```

figures 10
tables 0
includegraphics 1
labels 9
refs 7
floats 10

```

```

True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are referred to: (refs >= labels)

```

#### Label/ref check

```

119: In train.csv data file, it contains $42000$ gray-scale images of
hand-drawn digits, from zero through nine. Each image is a $28
\times 28$ pixels matrix with a total of 784 pixels
\cite{kaggle}. Each pixel has a single pixel-value which is an
integer from 0 to 255 associated with it, indicating the
lightness or darkness of that pixel (higher numbers meaning
lighter). In this experiment, we have plotted the graph in order
to see the appearance of these digits easily. Figure 1 shows the
first 70 samples.
364: From table 1, we can find that the Decision Tree algorithm has a
highest accuracy 0.8378 when we use Clean data. That's because
the Clean data contains all 784 features in the data set. It has
the minimum information loss among all three data set. Clean data

```

also have the longest running time, which is 20 seconds.

- 459: From table 2, we can find that Clean Data have a really low accuracy with the highest time spent. That's because the raw data set did not match the assumption of Naive Bayes. The features are not conditionally independent of each other. The pixels are continuous. For example, if pixel1 and pixel3 are both greater than 0, pixel2 will have a more probability to have a value greater than 0.
- 644: From table 4, we can find that Clean Data performed perfectly in this case. It takes the shortest time and reached a 0.9647 accuracy.
- 788: From table 6, we can easily find that when we use SVM classifier on PCA Data, we will receive the highest accuracy among all 5 different algorithms. The highest accuracy we reached for this project is 0.9813, which shows that our classifier predicts 98.13\% of the sample correct by using our SVM classifier. The time of training the model takes 127 seconds. The time spent is acceptable. The accuracy of SVM on PCA Clean Data has the second highest accuracy, which is 0.9785. The difference between first and second highest accuracy is about 0.0028, which is really small. However, the time spent saved 41.1\%. Therefore, SVM on PCA Clean Data is also a reasonable choice for the Digit Recognition task.

passed: False -> labels or refs used wrong

When using figures use columnwidth

[width=1.0\columnwidth]

do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

WARNING: figure and above may be used improperly

- 141: As we mentioned above, it can be seen from both the figure and the pixel-value that the value varies from 0 to 255, which means each feature is a continuous value. Thus, it is possible that such continuous values might affect our later feature selection. Our observation shows that the values are not very high at the

boundaries of 0 and \$>0\$. So here exist three ways to handle it  
\cite{data\_clean}:

WARNING: code and below may be used improperly

673: In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below Python code \cite{svm.code} in Figure \ref{fig:svm\_p}:

WARNING: algorithm and below may be used improperly

673: In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below Python code \cite{svm.code} in Figure \ref{fig:svm\_p}:

bibtex

---

label errors

89: feature\_extra: do not use underscore in labels:  
243: data\_clean: do not use underscore in labels:  
252: app1, title: do not use ' ' (spaces) in labels:  
263: app2, title: do not use ' ' (spaces) in labels:  
268: 99, title: do not use ' ' (spaces) in labels:  
278: second, title: do not use ' ' (spaces) in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--entry type for "cart" isn't style-file defined  
--line 5 of file report.bib  
Warning--entry type for "sklearn.cv" isn't style-file defined  
--line 57 of file report.bib  
Warning--entry type for "sklearn.dt" isn't style-file defined  
--line 68 of file report.bib  
Warning--entry type for "svm.form" isn't style-file defined  
--line 117 of file report.bib

```
Warning--entry type for "10f.cv" isn't style-file defined  
--line 129 of file report.bib  
Warning--entry type for "sp.rfc" isn't style-file defined  
--line 139 of file report.bib  
Warning--entry type for "nb.steps" isn't style-file defined  
--line 150 of file report.bib  
Warning--entry type for "svm.code" isn't style-file defined  
--line 162 of file report.bib  
Warning--entry type for "ss.dt" isn't style-file defined  
--line 174 of file report.bib  
Warning--entry type for "lr.form" isn't style-file defined  
--line 185 of file report.bib  
Warning--entry type for "wiki.nb" isn't style-file defined  
--line 208 of file report.bib  
Warning--entry type for "wiki.rf" isn't style-file defined  
--line 219 of file report.bib  
Warning--entry type for "wiki.svm" isn't style-file defined  
--line 231 of file report.bib  
Warning--no number and no volume in PCA  
Warning--page numbers missing in both pages and numpages fields in PCA  
Warning--no number and no volume in DT  
Warning--page numbers missing in both pages and numpages fields in DT  
Warning--no journal in app2  
Warning--no number and no volume in app2  
Warning--page numbers missing in both pages and numpages fields in app2  
Warning--no number and no volume in feature_extra  
Warning--page numbers missing in both pages and numpages fields in feature_extra  
Warning--no number and no volume in NB  
Warning--page numbers missing in both pages and numpages fields in NB  
Warning--no number and no volume in LR  
Warning--no number and no volume in hdwc  
Warning--page numbers missing in both pages and numpages fields in hdwc  
Warning--empty address in intro.dm  
Warning--no journal in data_clean  
Warning--no number and no volume in data_clean  
Warning--page numbers missing in both pages and numpages fields in data_clean  
(There were 31 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

-----  
passed: True

ascii  
-----

non ascii found 65292  
=====

The following tests are optional  
=====

Tip: newlines can often be replaced just by an empty line

find newline  
-----

passed: True

cites should have a space before \cite{} but not before the {  
-----

find cite {  
-----

passed: True

# Income Prediction based on Machine Learning Techniques

Borga Edionse Usifo

Indiana University

Bloomington, Indiana 47408

busifo@iu.edu

## ABSTRACT

This project takes a closer look to some of the most used supervised learning algorithms in machine learning. We start with the description of the each of the algorithms then we move it to analytics and findings by using that particular algorithm in our data-set. We also provide advantages and disadvantages of each supervised machine learning algorithm for future reference. We mainly focus on our prediction of the income level of individuals by looking at their age, gender, education, location, and other features given by our data-set. We will try each algorithm and try to pick the best features from our data-set to have an optimal prediction.

## KEYWORDS

i523, HID343, Machine Learning, Income Prediction, Logistic Regression, Ensemble methods

## 1 INTRODUCTION

In this project, we try to showcase the performance of the machine learning algorithms on data which we gather from UCI machine learning repository [22]. This data used by Kohavi R. and Becker B. for their research in improving the in Naive Bayes Classifier's accuracy [21].

Data consists of 15 variables, and we try to predict the income of the individuals. To do this prediction task, we first started with data preparation because the data we receive from UCI machine learning repository [22] not fully prepared for any machine learning algorithm. Our first task was the clean the data while applying some statistical techniques to get insights from the dataset. We also used data transformation methods like One-Hot-Encoding[45] to apply logarithmic functions for improving the machine learning algorithms performance before training the data.

Machine Learning algorithms that we discuss in this paper are Gaussian Naive Bayes [46], K Nearest Neighbors [29], Ensemble Methods (Boosting) [8], Support Vector Machines [6], Logistic Regression [34], and Decision Trees [49]. We try to show their weakness, advantages, and their time consumption while training each of them in machine learning algorithms section.

After providing a brief introduction of each of the supervised machine learning algorithms, we will discuss our findings for of each of the algorithms by comparing their accuracy score, F-1 score, recall, and lastly time comparison.

## 2 IMPORTANCE OF BIG DATA ANALYTICS FOR PREDICTIVE CLASSIFICATION

Importance of big data analytics is getting higher every day since the algorithms become more powerful to predict, classify and cluster any given data set. Importance of our case is any company can be used to predict individuals income to refer them goods in their

income range or governments can provide additional support for the areas that have lower income range. There can be many possible things that can do with this kind of classification predictions.

## 3 DATA PREPARATION

We first used the pandas [28] to help to load the data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame [48]. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was "income" attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables ( age, education, capital gain, capital loss, hours worked) by using the pandas [27] functions as shown in Table 1.

[Table 1 about here.]

### 3.1 Data Cleaning

After getting a snapshot from income data frame, we recognized that there is a column which has no meaning. The first task was to remove this entire column from our dataset we used pandas drop function for doing this task. After removing this column, we had more concise dataframe to analyze.

Moreover, removing the column we have encountered some missing values which labeled as "question marks" in data frame. In order to remove this values we first changed all the "question mark" values to "NaN" values by using pandas "replace" function [26]. After replacing all the question marks with "NaN" values, we used pandas missing value dropping function to remove all the "NaN" values from our dataset.

Furthermore, we start investigating the types of the variables, and in our case, we found two types of variable one of them labeled as "int64" which stands for integer values, other one labeled as object type of variable. From our previous example especially in "scikit-learn" it is better to use float object rather than "int64" for training the machine learning algorithms. Because their numerical output most of the time is "float64" object. We transferred all the "int64" objects to "float64" objects. This was the last step of the cleaning process.

Our last process is changing the string values to numerical values on our target data which consist of string values ("\$ 50K") for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table 2.

[Table 2 about here.]

Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called

one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well [45]. In other words “we implement binarization of the category to include as a future to train model [45]”. As we can see in Table 3 and Table 4.

[Table 3 about here.]

[Table 4 about here.]

## 4 DATA EXPLORATION

After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than \$50K and less than \$50K which can be seen in below Table 5.

[Table 5 about here.]

Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table 6. As we can see we have individuals who’re age ranging from 17 to 90 years old with a mean of 38.58. If we look at the capital gains and capital losses, we have a standard deviation of 7385 and 402 respectively this is also another indication of skew in these variables.

[Table 6 about here.]

We used scatter matrix plot and applied the correlation function to see if we have any reliable correlation between any of the variables. As we can see from the correlation matrix Table 7 and correlation numbers Figure 1 we do not have the high correlation between any variables. Correlation values range between -1 to 1. The correlation value of 1 is an indication of perfect positive correlation and correlation number -1 indicates a negative correlation between variables [15]. Because of lower correlation values, it will be tough to determine the classification by just looking at the correlations; this indicates we have sophisticated algorithms to determine the relationship between variables to classify individuals incomes.

[Table 7 about here.]

[Figure 1 about here.]

Furthermore, we also explore the capital gains, capital losses, and hours per week variables which we used a histogram to plot the data into distribution form so we can see how all these attributes distributed. The reason we do the histogram is we want to see any skewness in our data. As shown in the histogram graphs in Figure 2 and Figure 3 in capital gains and capital loss we have highly skewed data which can cause issues later on in our algorithms. We apply a logarithmic function to do highly skewed data to less skewed [24]. Using logarithmic functions adds more value to data from the interpretable standpoint and “it helps to meet the assumptions of inferential statistics [24]”.

[Figure 2 about here.]

[Figure 3 about here.]

Moreover, applying logarithmic function had an impact on distribution. We can see the changes on skew data in Figure 4 after applying logarithmic function.

[Figure 4 about here.]

## 5 MACHINE LEARNING ALGORITHMS TO CONSIDER

We have multiple algorithms to consider when we are doing the supervised learning. Each algorithm has its benefits and drawbacks. We will consider several supervised machine learning algorithms for our predictions. The application we will use to implement these algorithms will be Python Scikit-Learn library. We will briefly explain each parameter included in these algorithms in Scikit-Learn.

First we’ll look at the Scikit-Learn in Python framework we will go through the advantages in Scikit-Learn how we can implement any machine learning in just couple of simple line of codes in Scikit-Learn.

### 5.1 Why Scikit-Learn?

Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies[17]. “It also has a goal of providing common algorithms to Python users through consistent interface[2]”. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below[23]:

**Supervised Learning Algorithms:** One of the most fundamental reason that Scikit-Learn’s popularity comes from highly available supervised learning algorithms. These algorithms vary from regression models to decision trees and many more[23].

**Cross Validation:** Scikit-Learn includes various techniques to check the accuracy or any statistical measure between training and unseen testing set[23].

**Unsupervised Learning Algorithms:** Scikit-Learn had also various algorithms to support many unsupervised algorithms some of these include clustering, factor analysis, and neural network analysis[23].

**Various example data-sets:** Scikit-Learn comes with different data sets included in its package so users can start learning Scikit-Learn without the need of any data-sets[23].

**Feature extraction:** It has rich feature for extracting images or text from data-sets[23].

Algorithms that we will investigate shown below; we will go more deep analysis on each of these algorithms.

- Gaussian Naive Bayes
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Stochastic Gradient Descent Classifier
- Support Vector Machines
- Decision Trees

### 5.2 Gaussian Naive Bayes

Naive Bayes bring many beneficial features; it is widely popular among machine learning applications[41]. The popularity of Naive Bayes comes from being able to handle large projects and data-sets faster than most algorithms[41]. It also can handle complex data-sets with categorical and non-categorical inputs [41]. Naive Bayes based on probabilistic classifier of Bayesian theory. It is also a favorite way of doing text categorization [46].

Term naive comes from it is the method of use probability among categories which assumes of independence among given class of attributes as shown in Figure 5. In other words, if we try to classify individuals from their email communications it will not take the order of words into account. Whereas in the English language we can tell the difference between sentence makes sense or not if we randomly re-order our words in the sentences. So it does not understand the text, it only looks at word frequencies as a way to do the classification. This is why it is called “Naive”.

[Figure 5 about here.]

As we state above Naive Bayes derives from Bayesian Theory where the dimensionality of inputs is relatively high. Bayesian Theorem is stated below [16].

$$P(C | X) = \frac{P(X | C) \times P(C)}{P(X)} \quad (1)$$

Naive Bayes Classifier works as follows [16]:

#### **Advantages of Naive Bayes [16]:**

- Faster classification time for training data-set.
- Because of independent classification it improves classification performance.
- Performance is relatively good.

#### **Disadvantages of Naive Bayes[16]:**

- Often it requires a large number of data-sets to give adequate results.
- On some occasions which are relative to data-sets, it can give less accuracy.

### **5.3 Logistic Regression**

Logistic Regression widely used for predicting “probability of failure in a given system, product, and process [34]”. Logistic Regression also used in natural language analysis, it is an extension of conditional random fields [34]. It works as a classifier which learns the features from the input given and classifies them by multiplying the input value with the weight value [14].

$$P(C | X) = \sum_{i=1}^N W_i \times f_i \quad (2)$$

Main reason that Logistic Regression differs from Linear Regression is output variable for Logistic Regression is binary whereas output variable in Linear Regression is discrete(continuous) [12].

#### **Advantages of Logistic Regression:**

- It does not have any assumptions over distribution of classes [18].
- It is fast to train [18].
- Logistic Regression has fast classifying method of unknown data [18].
- We can easily extend to other regression for multiple classes like multinomial regression [18].

#### **Disadvantages of Logistic Regression:**

- One of the disadvantages of linear regression is it is not providing flexibility in some instances. What we mean by the “lack of flexibility is the linear dependency, and linear decision boundary in the instance space is not valid [42]”.

This disadvantage can be improved changing from Logistic Regression to Choquistic Regression[42].

- Logistic regression can provide poor results when there are more complex relationships in data [9].
- Logistic models also have over-fitting problems which come from a result of sampling bias [31].
- Because of Logistic Regression’s predictions comes from the independent variable if the researcher includes wrong independent variables then model’s prediction will have no value [31].
- Because it is predictions based on 1 and 0 model will have poor performance when predicting continuous variables [31].

### **5.4 K-Nearest Neighbors (KNN)**

K Nearest neighbor has been primarily studied, and this popularity comes from it has been applied to many applications some of these applications are “spatial databases, pattern recognition, geographic information, image retrieval, computer game, and many other applications [29]”. Due to an increase of mobile devices and people tends to use of applications like navigation K-nearest neighbor found itself another widely used area of location-based services due to an ability to found a target location [29].

Intuition behind the K Nearest Neighbor can be described as follows: “ for a set P of n objects and a querying point q, return the k objects in P that are closest to q [29].“

#### **Advantages of K Nearest Neighbors:**

- K Nearest Neighbor is a basic and simple approach to implement [35].
- K Nearest Neighbor can perform well and efficiently with the large amount of data [43].
- K nearest Neighbor also does effectively well with noisy data sets (“if the inverse square of weighted distance used as the distance [43]”). In other words, it is flexible to feature and distance choices [35].

#### **Disadvantages of K Nearest Neighbors:**

- K Nearest Neighbor typically require large dataset to perform well [35].
- Time complexity could be high due to computing distance of each query to all training data points [43]. This time might be improved with some indexing (K-D Tree) [43].
- Determining the value of K can be time-consuming [43].
- It can be unclear to know which type of distance to use, as well as which variability to use to get the optimal results [43].
- Switching the different K values can result in the predicted class labels [30].

Many of these disadvantages are improving with the help of parallel distributed computing. Recent improvements in MapReduce framework allows users to run KNN algorithms in the cluster which had a significant effect on reducing the computation time [19].

Another area of improvements on KNN, is to implement different mapping functions such as kernel KNN, kernel difference weighted KNN, adaptive quasi-conformal kernel nearest neighbor, angular

similarity, local linear discriminant analysis, and Dempster-Shafer [10].

## 5.5 Decision Trees

Decision Tree is another widely used algorithm model for classification and regression. Decision Trees uses a recursive split model where each recursive split is identified by each data point; this is an example of non-parametric hierarchical model [13].

Representation of decision trees is as follows; we sort the instances from root to leaf nodes, this sorting gives insights about the classification of the instance, every outcome descending from the root node corresponds to possible values for that variable [33]. We can classify an instance by starting from the root node and checking the attributes labeled on that node and moving down from that node based on attribute given attribute values [33] as shown in Figure 6.

[Figure 6 about here.]

### Advantages of Decision Trees:

- Decision Tree applications are easy to interpret and understand [32]. This ease comes from their schematic representation [32]. Interpretation between alternatives can be expressed with single numerical number which is the expected value (EV) [32].
- Decision Trees can handle noisy or incomplete data-sets [32]. In other words it requires little effort of data preparation because of its flexibility [7].
- It can handle both nominal and numerical variables [32].
- It can be modified easily whenever the new information is available [32].
- 

### Disadvantages of Decision Trees:

- Because of its use of divide and conquer method they can demonstrate good performance if there are few attributes exist when the attributes level goes into large number decision tree become more complex which will result in poor performance [32].
- Decision Trees are also susceptible to training set which can give a result of over-fitting [32]. In other words, it can believe the training set completely which will give an abysmal performance on testing set.
- ID3 and C4.5 decision tree algorithms require discrete values as input data.

## 5.6 Stochastic Gradient Descent Classifier (SGD)

Stochastic Gradient Descent recently got became more popular because of its large-scale learning ability in machine learning problems [11]. It is a useful and straightforward way approach of linear classifiers under convex problems which is Support Vector Machines or Conditional Random Fields [3]. The originality of SGD derives from “Stochastic Approximation” which is a work from Robinson and Monroe [5].

### Advantages of Stochastic Gradient Descent:

- One of the advantages of stochastic gradient descent is, it is easy to implement [38].

- Stochastic Gradient Descent is also efficient because of each step only relies on a single derivative which makes the computational cost  $1/n$  than normal gradient descent [37].

### Disadvantages of Stochastic Gradient Descent:

- Stochastic Gradient Descent can be required to have many iterations, and it also requires some hyper-parameters [38].
- Feature scaling is a practice which is used in the standardization of range of independent variables [47]. SGD also used this feature scaling technique and it can be sensitive to feature scaling [38].
- Another drawback of Stochastic Gradient Descent is while using GPU they are hard to parallelize or distributing them using computer clusters [25].

## 5.7 Support Vector Machines

Support Vector Machines is fallen under the classification methods in machine learning [6]. It is also a robust classification method that has been widely found itself an area ranging from pattern recognition to text analysis [6].

Fitting a boundary between data points is the principle of the support vector machines. This boundary divides the data points between classes, and each similar data point puts under the same class classification [6]. After training the support vector machines with training data-set, we only need to check whether the test data lies under the boundaries for testing set. Another thing to consider is after it creates the boundaries of the data remaining training data becomes obsolete because we only need the core set of points which supports the boundaries to classify the new data set. This core data points called “support vectors”. It is called vector because of each data point contains a row of observed data values for attributes [6].

[Figure 7 about here.]

Traditionally boundaries are called “hyperplanes” and it is used to describe boundaries in more than three dimensions because they are hard or sometimes impossible to visualize [7]. Figure 7. Optimality of hyperplane expressed as a linear function which requires maximum distance between the identified classes. It only considers a small number of training examples to build this hyperplane. SVM hyperplanes based on “separation of positive (+1) and negative (-1) with the largest margin [39]“.

One of the main characteristic of the machine learning is to generalize. In other words, we want to give a general idea that tends to fit any of our testing datasets optimally. Support vector machines are a perfect regarding generalizations because once the training data fitted by the support vector machines other than support vector data inside the training data becomes redundant which means that even with the small changes inside the data will not have a significant effect on general boundaries [6].

### Advantages of Support Vector Machines:

- Generalizes the data well with the help of boundaries. Which reduces the overfitting [6].
- Classification accuracy in basic support vector machine will yield a 95 percent accuracy with a default settings [6].
- SVM can deliver a unique solution, because of optimality solution is convex. This will give an advantage over Neural

Networks which has multiple solutions in local minima [1].

#### **Disadvantages of Support Vector Machines:**

- One common disadvantage of SVM, is the lack of transparency because of its non-parametric techniques [1].
- Another biggest disadvantage of SVM is it requires high algorithmic complexity and high level of memory for the large-scale implementations [39].
- According to Burges, biggest limitation of the SVM is in the choice of kernel [4].

## **5.8 Ensemble Methods**

Ensemble methods goes into classification algorithm category, they are learning algorithms which uses weighted vote for it is prediction methods, in other words, it is learning rules over a small subset of data then we combine these rules which we learn from the small subset of data to make predictions and/or classification on the testing data [8]. The originality of the Ensemble method comes from Bayesian averaging, but with the recent algorithms include “Bagging, error-correcting, and boosting [8]“.

Bagging refers to simply the looking at data-sets and dividing the data-set to it is small subsets then learning the rules of that particular small subset. Next step is combining each learned rule from subsets to apply to more significant data set. Combining method mostly done with averaging the learned rules. Bagging also does better on testing set than standard Linear Regression analysis and linear regression does better on training set especially in third order polynomial [8].

#### **Stacking**

Boosting is another method used in Ensemble Methods. The difference from bagging is in boosting we need to pick subsets or examples that we are not good at in other words hardest examples. Then we combine these learned rules with the weighted mean instead mean used in bagging method.

Boosting is little different then bagging.

#### **Advantages of Ensemble Methods:**

- Prediction of the ensemble methods is better than most of the algorithms because of the combining methods intuition makes the model less noisy [36].
- They are more stable than other algorithms. [36]

#### **Disadvantages of Ensemble Methods:**

- Over-fitting may cause some disadvantages for ensemble learning but bagging operation will reduce this overfitting [36].

## **6 FITTING DATA INTO MACHINE LEARNING ALGORITHMS**

In this section, we will show the techniques we used on the execution of the prepared data into machine learning algorithms. Before fitting the data into the machine learning algorithms, we split the data into two sets. These sets are the training set and the testing set. We do splitting because of gaining an access of the future data will most likely be hard before future occurs, and because of this fact, it is a good idea to test our model with a dataset which our model has not seen it [40].

We used scikit-learn for splitting data into train and test we saved 20% of data for testing purposes as shown in Table 8 .

[Table 8 about here.]

Furthermore, after splitting the data we put all of our training data into to each of the machine learning algorithm to get their prediction results. We also provided code at the beginning and the end of each algorithm to calculate their running time.

Before we move further we need to discuss critical characteristics of a machine learning algorithm. These are;

- Confusion Matrix
- Accuracy
- Recall
- F-1 Score
- Precision

**6.0.1 Confusion Matrix:** Confusion matrix develops from 4 key elements. These elements are true positive, true negative, false negative, and false positive. As shown in Figure 8 about the constructing a confusion matrix. If we want to build a confusion matrix by targeting individuals who are making more than \$50K our true positive, true negative, false positive, and false negative explained below.

[Figure 8 about here.]

**True Positive (TP):** We can explain true positive as if the individuals make more than \$50K and our model correctly classifies them as individuals who makes more than \$50K, then this individual is in higher income range, in this case, we call it a true positive [20].

**True Negative (TN):** Intuition of true negative is if an individual makes less than \$50K and our model correctly classifies them as individuals who makes less than \$50K, then this individual is in lower income range. We call this true negative [20].

**False Negative (FN):** When an individual makes less than \$50K and our model incorrectly classifies them in higher income range by making a mistake causes a false negative to happen [20].

**False Positive (FP):** When an individual is making more than \$50K and our model classifies them in lower income range by mistake. This is called false positive [20].

**6.0.2 Accuracy:** Accuracy answers the question of how good is the model is. In our case this question will be out of all the individuals, how many did the models classify the individuals correctly. The mathematical expression of the accuracy is the ratio between the number of correctly classified points and the number of total points. We can think that if we have high accuracy, our model is excellent, but this is only where we have identical false positive and false negative values in our dataset [20].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

**6.0.3 Precision.** Precision answers the questions of out of all the points predicted to be positive how many of them were actually positive? If we translate this question into our case, we will have out of all the individuals that we are classified as lower income how many were actually have lower income. Higher precision indicates that we have low false positive rate [20]. Mathematical expression of precision is;

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

**6.0.4 Recall (Sensitivity).** Recall answers the question of “out of the points that are labeled positive how many of them were correctly predicted is positive ? ”. If we translate this to our case, we will have “out of the points that are labeled higher income how many of them correctly predicted is in higher income range ? ”. Mathematical expression of the recall is;

$$Precision = \frac{TP}{TP + FN} \quad (5)$$

**6.0.5 F-1 Score.** The F-1 score is the idea of giving a decision by looking at only one score which will include precision, and recall scores. We cannot just take the average of precision and recall because if either of them is very low. We need a number to be low, even if the other one is not. This will lead us to look at the harmonic mean, and it works as follow. Let's say we have two numbers X and Y. X is smaller than Y, and we have the arithmetic mean, and it always lies between X and Y. It is a mathematical fact that the harmonic mean is always less than the arithmetic mean which is closer to the smaller number than to the higher number. Mathematical expression of F-1 score is;

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

## 6.1 Results

Now we can look at the results from each of the machine learning algorithm. Results also showed in Table 9 with the visualization of Figure 10. We can also see the running time of the each of the algorithm in Figure 9. Support Vector Machines is the winner for the highest running time for training the algorithm.

[Figure 9 about here.]

[Table 9 about here.]

**6.1.1 Naive Bayes.** As shown in the Figure 10 we have a comparison of several supervised machine learning algorithms on our dataset. We can see that from the accuracy standpoint Naive Bayes algorithms have the lowest score which means that it did not do a good job for labeling true positives regards to all data but it did a good job in precision standpoint while doing a bad classification from recall standpoint. Two key element for us in this situation is accuracy and f1 score(which consist of precision and recall).

**6.1.2 Support Vector Machine.** Support Vector Machine is the second best algorithm in our case. This algorithm did very well job on classification it has the second highest accuracy and f1 score.

**6.1.3 AdaBoost.** As we stated before ensemble algorithms learn from the small portion of the data and combine these learning to do the predictive task. As shown in Figure 10 adaboosting has the highest accuracy score among all the other algorithms. This algorithm should be our first choice to do predictive modeling. We believe that there is still an improvements on accuracy

**6.1.4 K-Nearest Neighbors.** K-Nearest Neighbor algorithm in our project we set the k value to 5. K Nearest Neighbor algorithm also did a good job by placing itself third in accuracy score.

**6.1.5 Decision Tree.** Decision Tree is gave a good accuracy but fall behind on f1 score as shown in Figure 10.

[Figure 10 about here.]

## 7 CONCLUSION

We presented the importance of analytical approach with machine learning algorithms and how they can be used to predict or classify the individuals with many different attributes like age, education, income, etc. We also presented weaknesses and strengths of these algorithms along with their precision, accuracy, recall, and F-1 scores by presenting with the visualizations. We also demonstrated the running time for each algorithm while using big data sets. The source code of this project can found Github website which presented in reference section [44].

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] L. Auria and A. R. Moro. 2008. Support Vector Machines (SVM) as a Technique for Solvency Analysis. Online. [http://www.diw.de/english/products/publications/discussion\\_papers/27539.html](http://www.diw.de/english/products/publications/discussion_papers/27539.html)
- [2] L. Ben. 2015. Six Reasons why I recommend scikit-learn. Online. (Oct. 2015). <https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn>
- [3] L. Bottou. 2010. Stochastic Gradient Descent. Online. (2010). <http://leon.bottou.org/projects/sgd>
- [4] C. J. C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 2 (01 Jun 1998), 121–167. <https://doi.org/10.1023/A:1009715923555>
- [5] N. Deanna, S. Nathan, and W. Rachel. 2016. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming* 155, 1 (01 Jan 2016), 549–573. <https://doi.org/10.1007/s10107-015-0864-7>
- [6] B. Deshpande. 2013. When do support vector machines trump other classification methods. Online. (Jan. 2013). <http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-other-classification-methods>
- [7] B. Deshpande. 2011. 4 key advantages of using decision trees for predictive analytics. Online. (July 2011). <http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics>
- [8] G. T. Dietterich. n.d. Ensemble Methods in Machine Learning. (n.d.). <http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf>
- [9] EliteDataScience. 2016. Modern Machine Learning Algorithms: Strengths and Weaknesses. Online. (May 2016). <https://elitedatascience.com/machine-learning-algorithms>
- [10] O. F. Ertugrul and M. E. Tagluk. 2017. A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing* 55, Supplement C (2017), 480 – 490. <https://doi.org/10.1016/j.asoc.2017.02.020>
- [11] M. Fan. n.d. How and Why to Use Stochastic Gradient Descent? (n.d.). <http://anson.ucdavis.edu/~minjay/SGD.pdf>
- [12] J. Fang. 2013. Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses. *Journal of Business and Economics* 4, 7 (July 2013), 620–633. <http://www.academicstar.us/UploadFile/Picture/2014-6/201461494819669.pdf>
- [13] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem. 2017. Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renewable Energy* 111, Supplement C (2017), 52 – 62. <https://doi.org/10.1016/j.renene.2017.03.083>
- [14] S. T. Indra, L. Wikarsa, and R. Turang. 2016. Using logistic regression method to classify tweets into the selected topics. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on* 1, 385–389 (2016), 385. [http://proxyiub.uits.iu.edu/login.aspx?direct=true&db=edsee&AN=edsee.7872727&site=eds-live&scope=site](http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edsee&AN=edsee.7872727&site=eds-live&scope=site)
- [15] Investopedia. n.d. Correlation Coefficient. Online. (n.d.). <https://www.investopedia.com/terms/c/correlationcoefficient.asp>
- [16] D. S. Jadhav and H. P. Channe. 2014. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)* 5, 1 (Jan. 2014), 1842–1845. <https://www.ijsr.net/archive/v5i1/NOV153131.pdf>

- [17] B. Jason. 2014. A gentle introduction to Scikit-Learn: Python Machine Learning Library. Online. (April 2014). <https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/>
- [18] H. Jeff. 2012. Introduction to Machine Learning. Online. (Jan. 2012). [http://courses.washington.edu/css490/2012/Winter/lecture\\_slides/05b\\_logistic\\_regression.pdf](http://courses.washington.edu/css490/2012/Winter/lecture_slides/05b_logistic_regression.pdf)
- [19] J. Jiaqi and Y. Chung. 2017. Research on K nearest neighbor join for big data. In *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE, Department of Computer Engineering Wonkwang University Iksan 54538, Korean, 1077–1081. <https://doi.org/10.1109/ICInFA.2017.8079062>
- [20] R. Joshi. 2016. Accuracy, Precision, Recall, and F1 Score: Interpretation of Performance Measures. Online. (Sept. 2016). <http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures>
- [21] R. Kohavi. 1996. Improving the Accuracy of Naive-Bayes Classifiers: A Decision-tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Silicon Graphics, Inc, 202–207. <http://dl.acm.org/citation.cfm?id=3001460.3001502>
- [22] R. Kohavi and B. Becker. n.d.. Predicting whether income exceeds \$50K/yr based on census data. Online. (n.d.). <https://archive.ics.uci.edu/ml/datasets/Census+Income>
- [23] J. Kunal. 2015. Scikit-Learn in python - The most important Machine Learnig Tool I learnt last year. Online. (Jan. 2015). <https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/>
- [24] M. D. Lane. n.d.. Log Transformations. Online. (n.d.). <http://onlinestatbook.com/2/transformations/log.html>
- [25] V. Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng. 2011. On optimization methods for deep learning. In *International Conference of Machine Learning*. Stanford University, International Conferenfe of Machine Learning, Stanford University, NA. <https://cs.stanford.edu/~acoates/papers/LeNgiCoaLahProNg11.pdf>
- [26] Pandas Library. n.d.. Dataframe replace. Online. (n.d.). <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html>
- [27] Pandas Library. n.d.. Pandas Dateframe describe. Online. (n.d.). <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html>
- [28] Pandas Py Data Library. n.d.. Pandas for Python. Online. (n.d.). <https://pandas.pydata.org/>
- [29] L. J. Moon. 2017. Fast k-Nearest Neighbor Searching in Static Objects. *Wireless Personal Communications* 93, 1 (01 Mar 2017), 147–160. <https://doi.org/10.1007/s11277-016-3524-1>
- [30] G. Nick. 2014. KNN. Online. (April 2014). <http://www.nickgillian.com/wiki/pmwiki.php/GRT/KNN>
- [31] R. Nick. NA. The Disadvantages of Logistic Regression. Online. (NA). <http://classroom.synonym.com/disadvantages-logistic-regression-8574447.html>
- [32] C. Petri. 2010. Decison Trees. Online. (2010). <http://www.cs.ubbcluj.ro/~gabis/DocDiplome/DT/DecisionTrees.pdf>
- [33] U. Princeton. NA. Decision Tree Learning. Online. (NA). <http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-decrees.pdf>
- [34] S. A. Raj, L. J. Fernando, and S. Raj. 2017. Predictive Analytics On Political Data. Congress. *World Congress on Computing and Communication Technologies* 10, 1109 (2017), 93–96.
- [35] M. Ray. 2012. Nearest Neighbours: Pros and Cons. Online. (April 2012). <http://www2.cs.man.ac.uk/~raym8/comp37212/main/node264.html>
- [36] S. Ray. 2015. 5 Easy Questions on Ensemble Modeling Everyone Should Know. Online. (Jan. 2015). <https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/>
- [37] J. Rie and Z. Tong. 2013. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Rutgers University, New Jersey, USA, 315–323. <http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf>
- [38] Scikitlearn. n.d.. Stochastic Gradient Descent. Online. (n.d.).
- [39] K. N. Shrivastava, P. Saurabh, and B. Verma. 2011. An Efficient Approach Parallel Support Vector Machine for Classification of Diabetes Dataset. *International Journal of Computer Applications in Technology* 36, 6 (Dec. 2011), 19–24. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.3757&rep=rep1&type=pdf>
- [40] D. Steinberg. 2014. Why Data Scientist Split Data into Train and Test. Online. (March 2014). <https://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test>
- [41] K. B. Tapan. 2015. Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *Inteligencia Artificial, Vol 18, Iss 56, Pp 14-30 (2015) 1, 56 (2015), 14.* <http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edsdjo&AN=edsdjo.0e372b34c5d48bcb72cd437eede1fd1&site=eds-live&scope=site>
- [42] A. F. Tehrani, W. Cheng, and E. Hullermeier. 2011. Choquistic Regression: Generalizing Logistic Regression Using the Choquet Integral. Online. (July 2011). <https://www-old.cs.uni-paderborn.de/fileadmin/Informatik/eim-i-is/PDFs/Talk.EUSFLAT.11.pdf>

#### LIST OF FIGURES

1	Scatter Matrix Plot [44].	9
2	Histogram of Capital Gain [44].	10
3	Histogram of Capital Loss [44].	10
4	After Logarithmic Function Applied Histogram of Capital Gain [44].	11
5	Example of Naive Bayes [50].	12
6	Example of Decision Tree Construction[33].	12
7	Example of Shows the Hyperplanes [6].	13
8	Example of Confusion Matrix Construction [20].	13
9	Supervised Learning Algorithm Running Time Results [44].	14
10	Supervised Learning Algorithm Results [44].	15

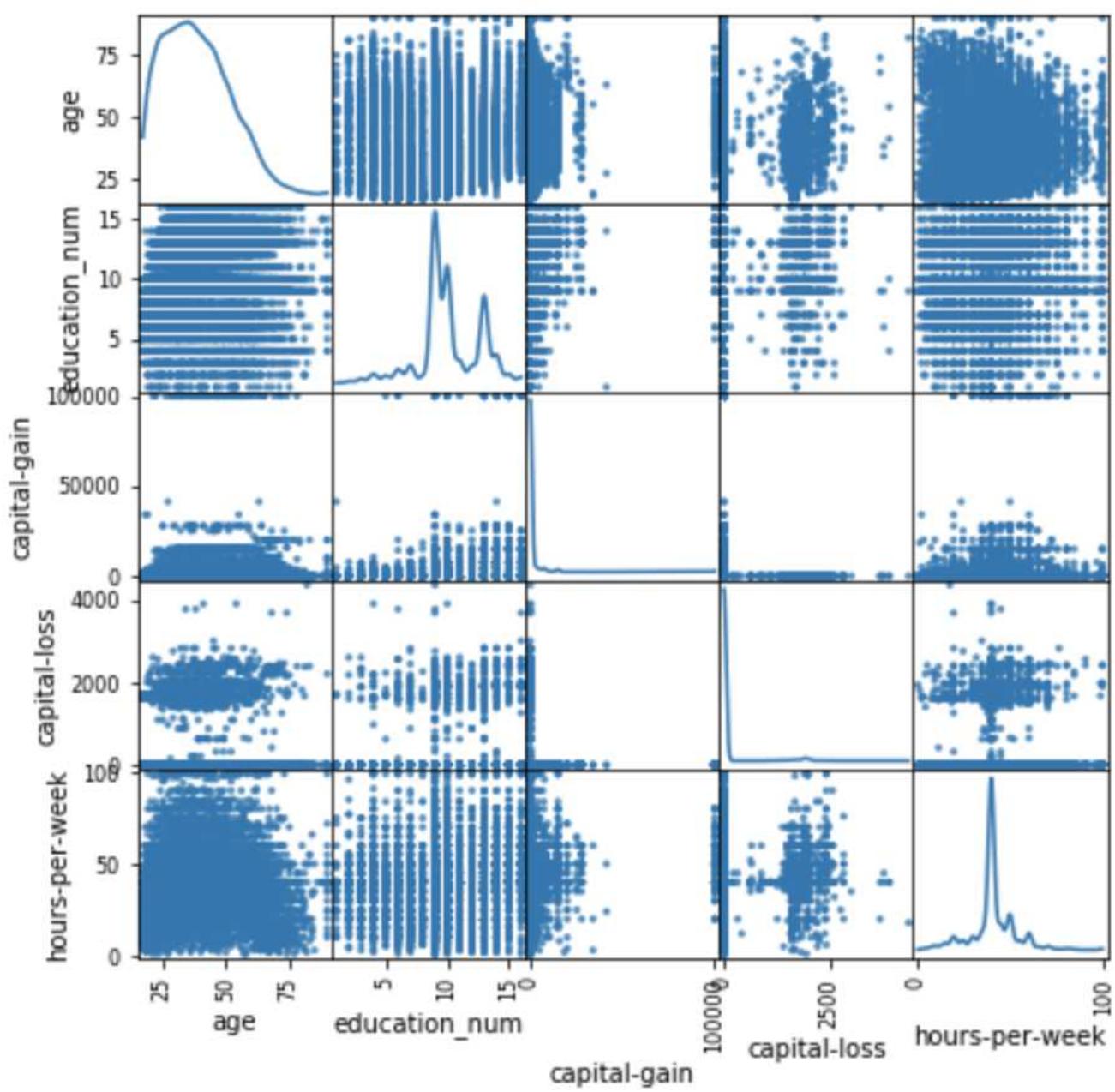
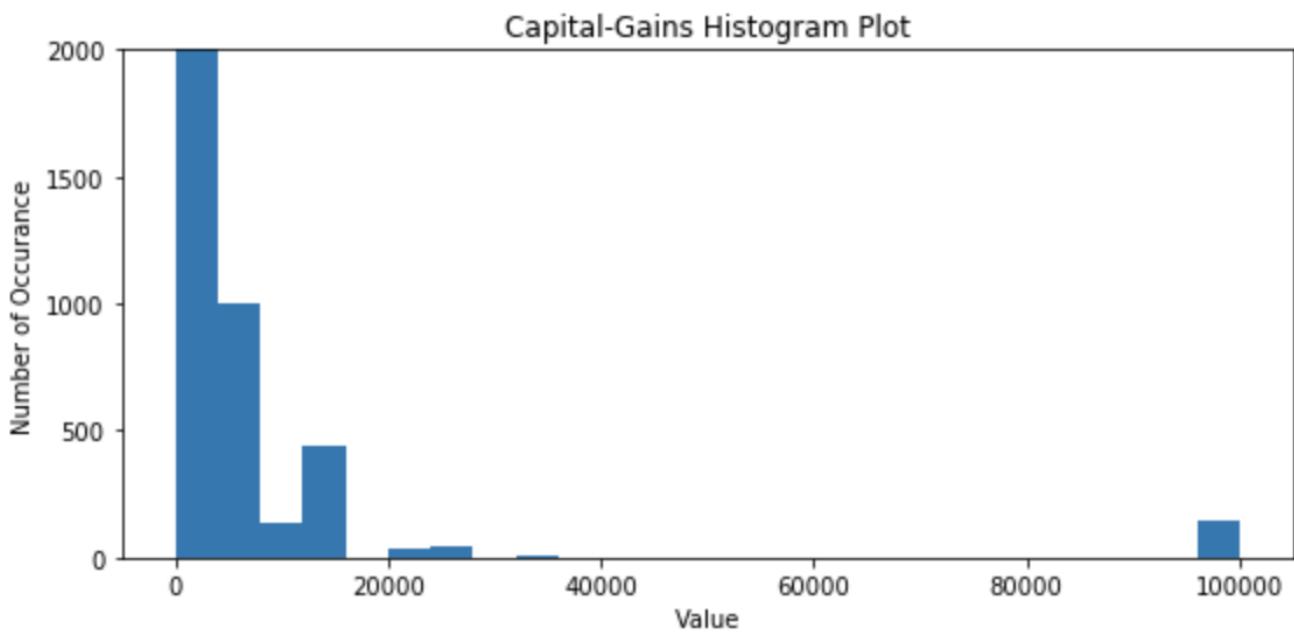


Figure 1: Scatter Matrix Plot [44].



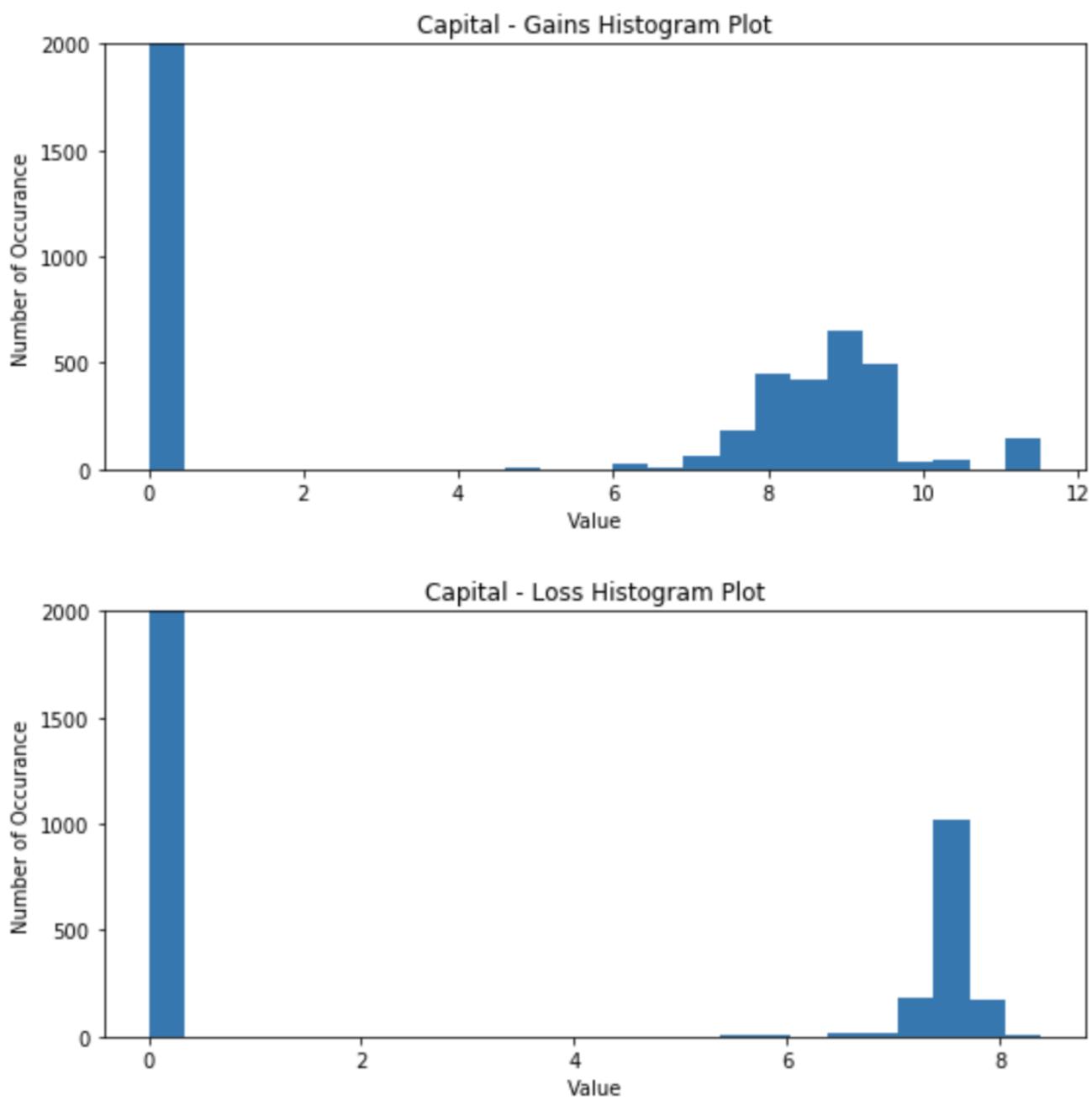


Figure 4: After Logarithmic Function Applied Histogram of Capital Gain [44].

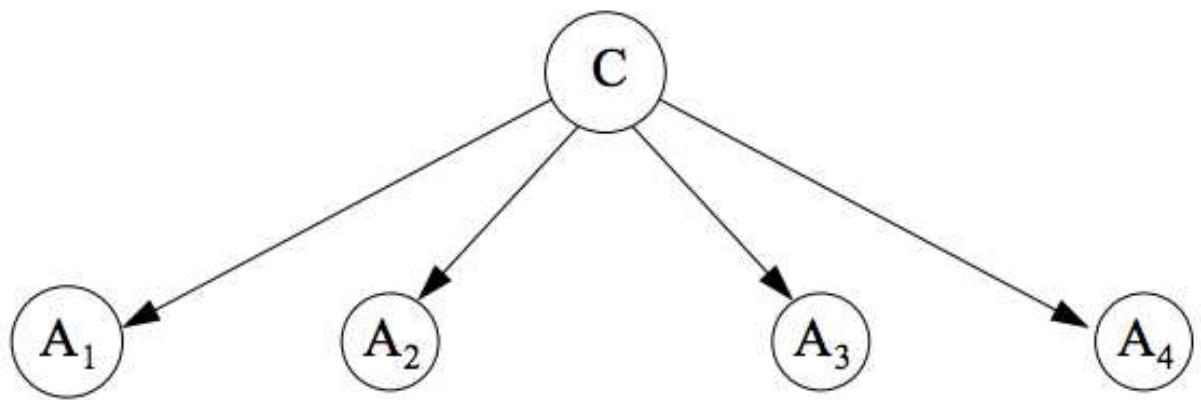


Figure 5: Example of Naive Bayes [50].

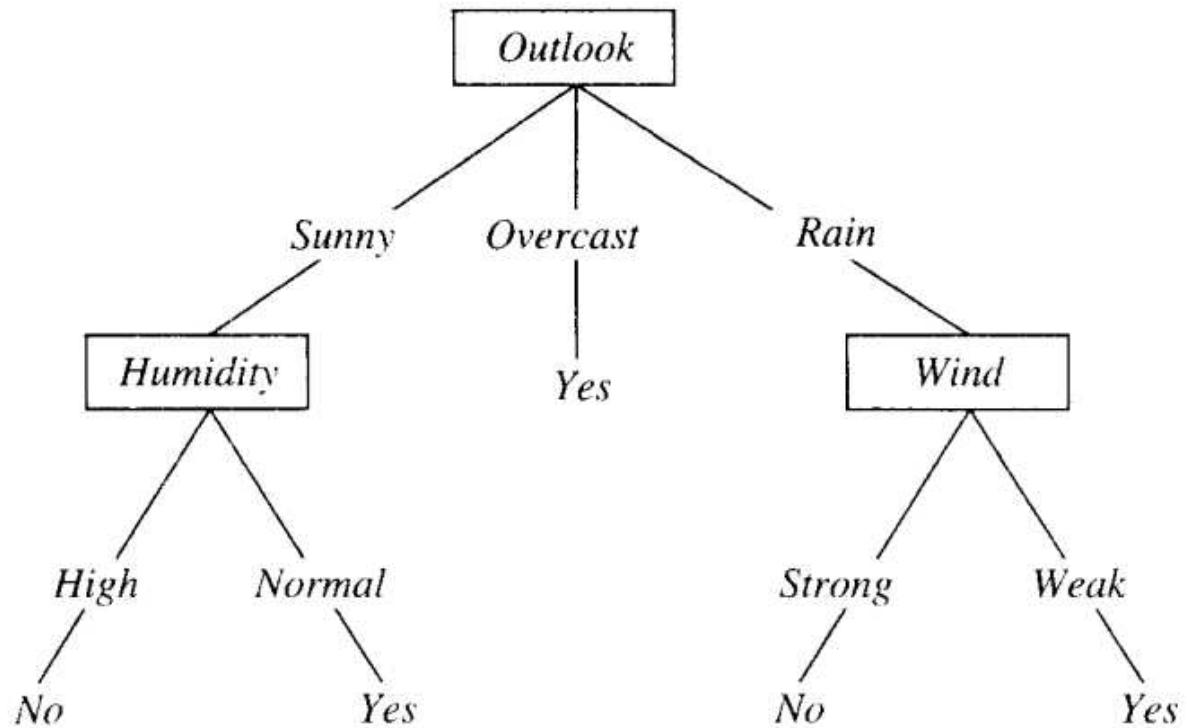


Figure 6: Example of Decision Tree Construction[33].

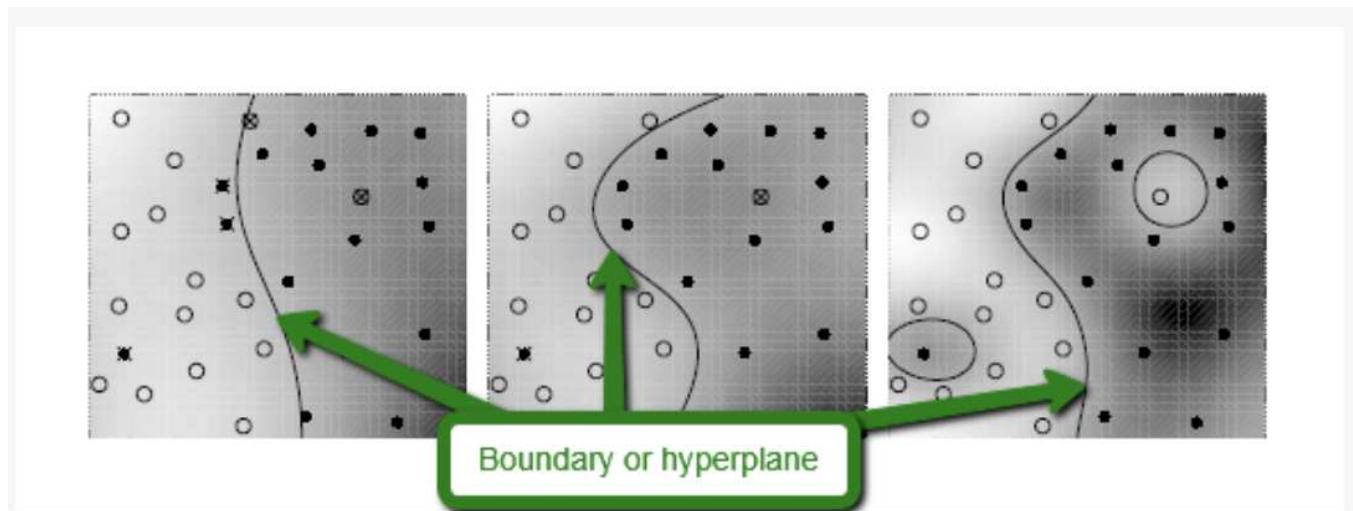
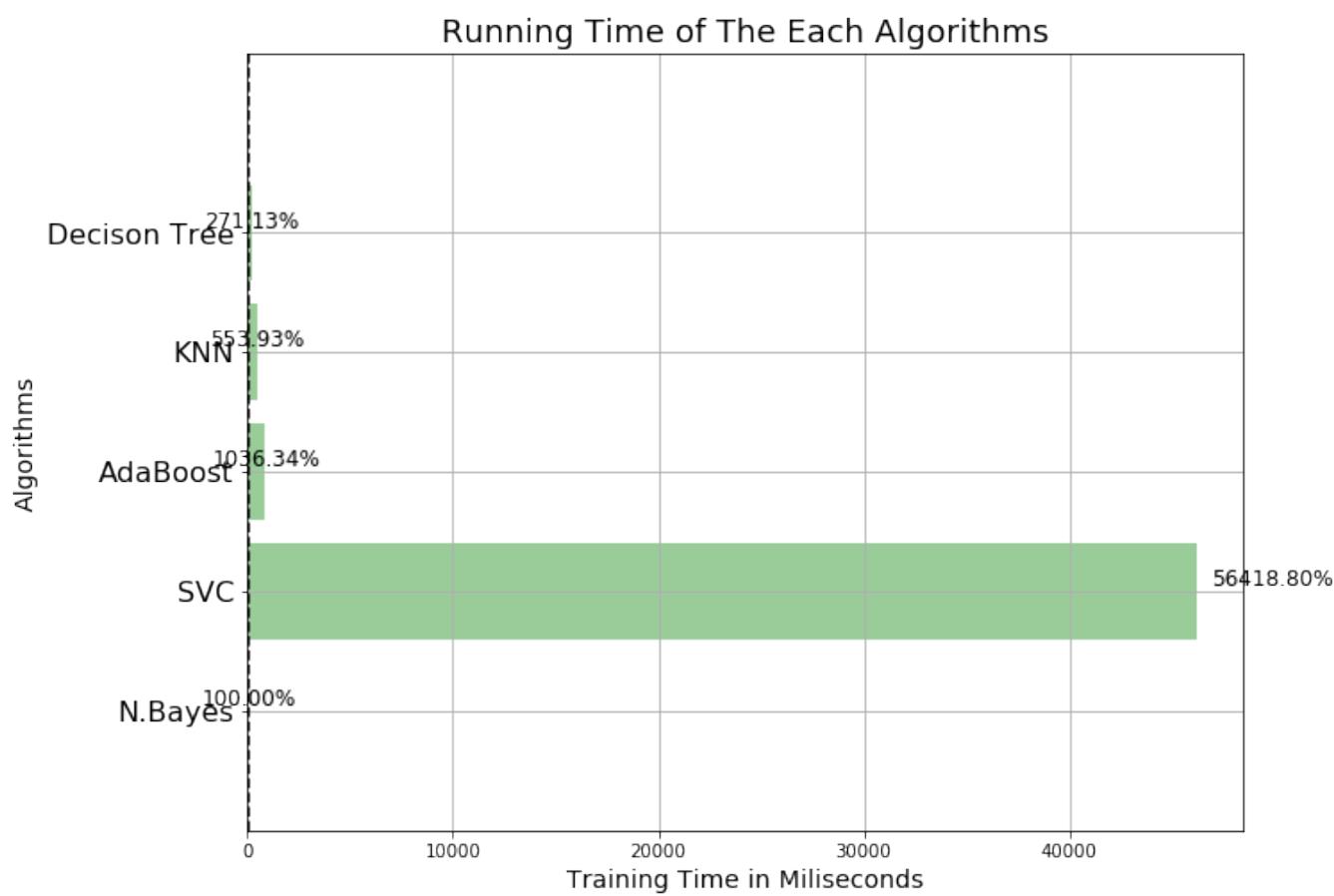


Figure 7: Example of Shows the Hyperplanes [6].

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 8: Example of Confusion Matrix Construction [20].



**Figure 9: Supervised Learning Algorithm Running Time Results [44].**

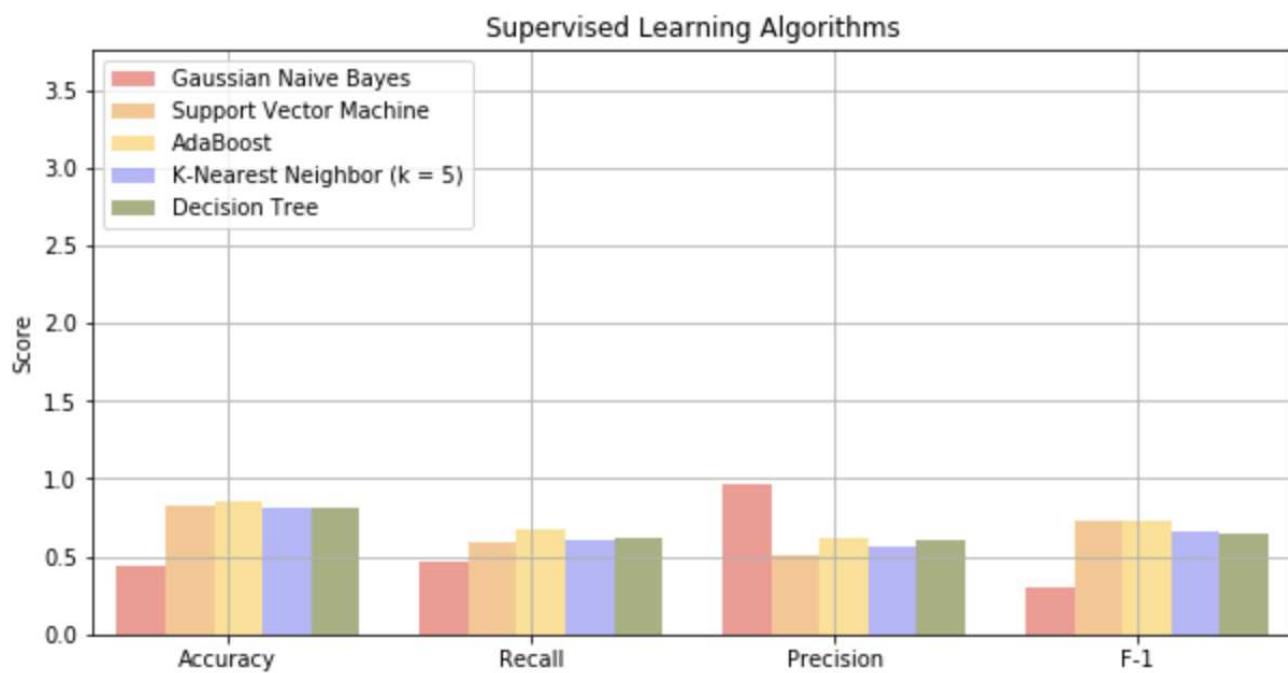


Figure 10: Supervised Learning Algorithm Results [44].

#### LIST OF TABLES

1	Statistical Summary of The Continuous Variables	17
2	Description of the Binary Values	17
3	Example of One Hot Encoding Before [45].	17
4	Example of One Hot Encoding After [45].	17
5	Count of Income Variable Regarding to Individuals	17
6	Statistical Summary of Continuous Variables [44].	18
7	Correlation Matrix [44].	18
8	Train-Test-Split [44].	18
9	Results of the Algorithms [44].	18

	<b>age</b>	<b>education</b>	<b>cap gain</b>	<b>cap loss</b>	<b>hours</b>
<b>count</b>	32561	32561	32561	32561	32561
<b>mean</b>	38.581	10.08	1077.64	87.303	40.437
<b>std.</b>	13.640	2.572	7385.292	402.960	12.347
<b>min.</b>	17.0	1.0	0	0	1.0
<b>25%</b>	28.0	9.0	0	0	40.0
<b>50%</b>	37.0	10.0	0	0	40.0
<b>75%</b>	48.0	12.0	0	0	45.0
<b>max</b>	90.0	16.0	0	4356.0	99.0

**Table 1: Statistical Summary of The Continuous Variables**

Description	Assigned Value
Individuals who makes more than \$50K	1
Individuals who makes at or less than \$50K	0

**Table 2: Description of the Binary Values**

<b>Company Name</b>	<b>Categorical Variable</b>	<b>Price</b>
VW	1	2000
Acura	2	10011
Honda	3	50000
Honda	3	10000

**Table 3: Example of One Hot Encoding Before [45].**

VW	Acura	Honda	Price
1	0	0	20,000
0	1	0	10,011
0	0	1	50,000
0	0	1	10,000

**Table 4: Example of One Hot Encoding After [45].**

<b>Description</b>	<b>Count</b>
Total Number of Individuals	30162
Individuals who makes more than \$50K	7508
Individuals who makes at or less than \$50K	22654

**Table 5: Count of Income Variable Regarding to Individuals**

	<b>Age</b>	<b>Gain</b>	<b>Loss</b>	<b>Hours</b>
<b>Number of Instances</b>	32,561	32,561	32,561	32,561
<b>Mean</b>	38.58	1077.64	87.303	40.437
<b>Standard Deviation</b>	13.640	7385.292	402.960	12.347
<b>Minimum Value</b>	17	0	0	1
<b>25th percentile</b>	28	0	0	40
<b>50th percentile</b>	37	0	0	40
<b>75th percentile</b>	48	0	0	45
<b>Maximum Values</b>	90	99999	4356	99

Table 6: Statistical Summary of Continuous Variables [44].

	<b>Age</b>	<b>Education</b>	<b>Capital Gain</b>	<b>Capital Loss</b>	<b>Hours Per Week</b>
<b>Age</b>	1.0	0.043	0.080	0.060	0.101
<b>Education</b>	0.043	1.0	0.124	0.079	0.152
<b>Capital Gain</b>	0.080	0.124	1.0	-0.032	0.080
<b>Capital Loss</b>	0.060	0.796	-0.032	1.0	0.052
<b>Hours Per Week</b>	0.101	0.152	0.080	0.052	1.0

Table 7: Correlation Matrix [44].

<b>Splitting the Data</b>	<b>Sample Size</b>
Training	24129
Testing	6033

Table 8: Train-Test-Split [44].

Name	Accuracy	Recall	Precision	F1 Score
Naive Bayes	0.4442	0.4642	0.9680	0.3053
SVC	0.8301	0.5969	0.5056	0.7284
AdaBoost	0.8499	0.6724	0.6189	0.7361
KNN	0.8184	0.6090	0.5682	0.6561
Decision Tree	0.8161	0.6231	0.6109	0.6459

Table 9: Results of the Algorithms [44].

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
=====
bibtext _ label error
```

```
=====
report.bib:518:@incollection{NIPS2013_4937,
```

```
=====
bibtext space label error
```

```
=====
report.bib:172:@INPROCEEDINGS{knn-chung,
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-16 09.39.07] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.9s.
```

```
./README.yml
```

```
8:81      error    line too long (188 > 80 characters)  (line-length)
8:188     error    trailing spaces  (trailing-spaces)
21:81     error    line too long (534 > 80 characters)  (line-length)
34:81     error    line too long (666 > 80 characters)  (line-length)
34:666    error    trailing spaces  (trailing-spaces)
35:12     error    trailing spaces  (trailing-spaces)
37:30     error    trailing spaces  (trailing-spaces)
42:5      error    duplication of key "type" in mapping  (key-duplicates)
```

```
=====
Compliance Report
```

```
name: Usifo, Borga
hid: 343
paper1: 100 %
paper2: 100 %
project: 100 %
```

```
yamlcheck
```

---

```
wordcount
```

---

```
18
wc 343 project 18 5795 report.tex
wc 343 project 18 6330 report.pdf
wc 343 project 18 3252 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
49: We first used the pandas \cite{www-pandas} to help to load the
```

data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame \cite{www-commasep}. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was “income” attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables ( age, education, capital gain, capital loss, hours worked) by using the pandas \cite{www-pandas.describe} functions as shown in Table \ref{stats-table}.

```

51: \begin{table}[!ht]
66: \label{stats-table}
78: \par Our last process is changing the string values to numerical values on our target data which consist of string values (''\$ 50K'') for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table \ref{assign-values}.
80: \begin{table}[!ht]
89: \label{assign-values}
92: \par Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well \cite{www-hackernoon}. In other words ‘‘we implement binarization of the category to include as a future to train model \cite{www-hackernoon}’’. As we can see in Table \ref{one-hot-before} and Table \ref{one-hot-after}.
95: \begin{table}[!ht]
106: \label{one-hot-before}
110: \begin{table}[!ht]
121: \label{one-hot-after}
127: After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than \$50K and less than \$50K which can be seen in below Table \ref{my-label-2}.
129: \begin{table}[!ht]
139: \label{my-label-2}
143: \par Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table \ref{my-label}. As we can see we have individuals who’re age ranging from 17 to 90 years old with a mean of 38.58. If we look
  
```

at the capital gains and capital losses, we have a standard deviation of 7385 and 402 respectively this is also another indication of skew in these variables.

```
145: \begin{table}![ht]
160: \label{my-label}
163: \par We used scatter matrix plot and applied the correlation function to see if we have any reliable correlation between any of the variables. As we can see from the correlation matrix Table \ref{scatter-matrix} and correlation numbers Figure \ref{fig:scatter} we do not have the high correlation between any variables. Correlation values range between -1 to 1. The correlation value of 1 is an indication of perfect positive correlation and correlation number -1 indicates a negative correlation between variables \cite{www-investopedia}. Because of lower correlation values, it will be tough to determine the classification by just looking at the correlations; this indicates we have sophisticated algorithms to determine the relationship between variables to classify individuals incomes.
165: \begin{table}![ht]
178: \label{scatter-matrix}
182: \begin{figure}![ht]
184: \includegraphics[width=\columnwidth]{images/scatter-matrix.png}
185: \caption{Scatter Matrix Plot
\cite{Borga2017}.}\label{fig:scatter}
188: \par Furthermore, we also explore the capital gains, capital losses, and hours per week variables which we used a histogram to plot the data into distribution form so we can see how all these attributes distributed. The reason we do the histogram is we want to see any skewness in our data. As shown in the histogram graphs in Figure \ref{fig:Hist-capital} and Figure \ref{fig:loss-capital} in capital gains and capital loss we have highly skewed data which can cause issues later on in our algorithms. We apply a logarithmic function to do highly skewed data to less skewed \cite{www-onlinestat}. Using logarithmic functions adds more value to data from the interpretable standpoint and ‘‘it helps to meet the assumptions of inferential statistics \cite{www-onlinestat}’’.
190: \begin{figure}![ht]
192: \includegraphics[width=\columnwidth]{images/capital-gain.png}
193: \caption{Histogram of Capital Gain
\cite{Borga2017}.}\label{fig:Hist-capital}
196: \begin{figure}![ht]
198: \includegraphics[width=\columnwidth]{images/capital-loss.png}
199: \caption{Histogram of Capital Loss
\cite{Borga2017}.}\label{fig:loss-capital}
202: \par Moreover, applying logarithmic function had an impact on
```

distribution. We can see the changes on skew data in Figure \ref{fig:Hist-capital-log} after applying logarithmic function.

204: \begin{figure}[!ht]  
206: \includegraphics[width=\columnwidth]{images/logarithmic-applied.png}  
207: \caption{After Logarithmic Function Applied Histogram of Capital Gain \cite{Borga2017}.}\label{fig:Hist-capital-log}

244: \par Term naive comes from it is the method of use probability among categories which assumes of independence among given class of attributes as shown in Figure \ref{fig:Naive Bayes}. In other words, if we try to classify individuals from their email communications it will not take the order of words into account. Whereas in the English language we can tell the difference between sentence makes sense or not if we randomly re-order our words in the sentences. So it does not understand the text, it only looks at word frequencies as a way to do the classification. This is why it is called ‘‘Naive’’.

246: \begin{figure}[!ht]  
249: \includegraphics[width=\columnwidth]{Naive-bayes}  
250: \caption{Example of Naive Bayes \cite{Zhang}.}\label{fig:Naive Bayes}

335: \par Representation of decision trees is as follows; we sort the instances from root to leaf nodes, this sorting gives insights about the classification of the instance, every outcome descending from the root node corresponds to possible values for that variable \cite{www-cs.princeton}. We can classify an instance by starting from the root node and checking the attributes labeled on that node and moving down from that node based on attribute given attribute values \cite{www-cs.princeton} as shown in Figure \ref{fig:Decision Tree}.

337: \begin{figure}[!ht]  
339: \includegraphics[width=\columnwidth]{images/decison\_tree.png}  
340: \caption{Example of Decision Tree Construction\cite{www-cs.princeton}.}\label{fig:Decision Tree}

388: \begin{figure}[!ht]  
390: \includegraphics[width=\columnwidth]{images/hyperplane-boundary.png}  
391: \caption{Example of Shows the Hyperplanes \cite{www-simafore-svm}.}\label{fig:Hyperplane}

394: \par Traditionally boundaries are called ‘‘hyperplanes’’ and it is used to describe boundaries in more than three dimensions because they are hard or sometimes impossible to visualize.\cite{www-simafore}. Figure \ref{fig:Hyperplane}. Optimality of hyperplane expressed as a linear function which requires maximum distance between the identified classes. It only considers a small number of training example to build this

hyperplane. SVM hyperplanes based on “ separation of positive (+1) and negative (-1) with the largest margin \cite{verma-ssv}”.

439: \par We used scikit-learn for splitting data into train and test we saved 20\% of data for testing purposes as shown in Table \ref{split} .

441: \begin{table}[!ht]

450: \label{split}

466: Confusion matrix develops from 4 key elements. These elements are true positive, true negative, false negative, and false positive. As shown in Figure \ref{fig:confusion-matrix} about the constructing a confusion matrix. If we want to build a confusion matrix by targeting individuals who are making more than \\$50K our true positive, true negative, false positive, and false negative explained below.

468: \begin{figure}[!ht]

470: \includegraphics[width=\columnwidth]{images/confusion-matrix.png}

471: \caption{Example of Confusion Matrix Construction \cite{www-exsilio}.}\label{fig:confusion-matrix}

513: Now we can look at the results from each of the machine learning algorithm. Results also showed in Table \ref{result-table} with the visualization of Figure \ref{fig:result-algo}. We can also see the running time of the each of the algorithm in Figure \ref{fig:result-time}. Support Vector Machines is the winner for the highest running time for training the algorithm.

515: \begin{figure}[!ht]

517: \includegraphics[width=\columnwidth]{images/running-time.png}

518: \caption{Supervised Learning Algorithm Running Time Results \cite{Borga2017}.}\label{fig:result-time}

521: \begin{table}[!ht]

533: \label{result-table}

537: As shown in the Figure \ref{fig:result-algo} we have a comparison of several supervised machine learning algorithms on our dataset. We can see that from the accuracy standpoint Naive Bayes algorithms have the lowest score which means that it did not do a good job for labeling true positives regards to all data but it did a good job in precision standpoint while doing a bad classification from recall standpoint. Two key element for us in this situation is accuracy and f1 score(which consist of precision and recall).

541: As we stated before ensemble algorithms learn from the small portion of the data and combine these learning to do the predictive task. As shown in Figure \ref{fig:result-algo} adaboosting has the highest accuracy score among all the other algorithms. This algorithm should be our first choice to do predictive modeling. We believe that there is still an

```
improvements on accuracy
546: Decision Tree is gave a good accuracy but fall behind on f1 score
      as shown in Figure \ref{fig:result-algo}.
553: \begin{figure}[!ht]
555: \includegraphics[width=\columnwidth]{images/result-score.png}
556: \caption{Supervised Learning Algorithm Results
      \cite{Borga2017}.}\label{fig:result-algo}
```

```
figures 10
tables 9
\includegraphics 10
labels 19
refs 17
floats 19
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= \includegraphics)
False : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

```
WARNING: algorithm and below may be used improperly
```

```
127: After cleaning the data, we started our data exploration to learn
      little bit more from our data and make necessary changes if
      needed before putting into our machine learning algorithms. The
      first step in this process is getting the total count of the
      individuals as well as the count of the individuals who are
      making more than \$50K and less than \$50K which can be seen in
      below Table \ref{my-label-2}.
```

WARNING: code and below may be used improperly

220: Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies\cite{www-machinelearningmystery}. ‘‘It also has a goal of providing common algorithms to Python users through consistent interface\cite{www-oreily}’’. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below\cite{www-analyticvidhya}:

WARNING: algorithm and below may be used improperly

220: Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies\cite{www-machinelearningmystery}. ‘‘It also has a goal of providing common algorithms to Python users through consistent interface\cite{www-oreily}’’. Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below\cite{www-analyticvidhya}:

WARNING: algorithm and below may be used improperly

230: \par Algorithms that we will investigate shown below; we will go more deep analysis on each of these algorithms.

bibtex

---

label errors

518: NIPS2013\_4937: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst  
Database file #1: report.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Big Data Analytics on Influencers in Social Networks

Bharat Mallala  
Indiana University  
Bloomington, IN 47408, USA  
bmallala@iu.edu

Jyothi Pranavi Devineni  
Indiana University  
Bloomington, IN 47408, USA  
jyodevin@iu.edu

## ABSTRACT

Social Networking has become a part of people's lives with millions of posts made every day. This huge volume of data paves the way to find interesting insights from data. Applying Big Data analytic methods on Twitter data helps in identifying the most influenced users. The Twitter data used for analysis is taken from Kaggle website. Performing feature selection tasks on this data has helped in identifying the important features that differentiate a user's influence over the other. By applying various Machine learning classification algorithms iteratively on these features obtained can be used to classify a user as more influential over the other.

## KEYWORDS

I523, hid215, hid208, Machine Learning, Artificial Intelligence, Data Analysis, Multi layer Neural Network, Logistic regression, Random Forest Classifier, Support Vector Classifier(SVC), Stochastic Gradient Descent(SGD), K Nearest Neighbours(KNN), Naive Bayes.

## 1 INTRODUCTION

Social Media is a powerful tool to express one's thoughts, ideas and experiences. With the exponential growth in active users across various social media platforms such as Facebook, Twitter, Instagram etc, a mammoth amount of data is being generated across these platforms. With such huge volume of data readily available, a lot of research is being carried on in the recent time in analyzing this data using various tools and extracting useful insights from it. With such large number of active users on these platforms, organizations are looking to capitalize on the opportunity to reach out to a large mass of population with absolutely low marketing costs.

Post made on Twitter are short and concise making it one the most used social media platform used by millions all around the world. Contrary to other social media platforms, Twitter has a limit on the number of words used by its users to make posts, making it the most professional way to interact with people. Most influenced people around the world such as politicians, sportsmen, businessmen, artists are active users of Twitter. With such large number of celebrities using Twitter to express themselves, it is interesting to find how influential they are when compared to their counterparts. One can argue that a user's influence on Twitter is directly proportional to the number of followers he/she has. This is debatable as a set of people argue that with the number of followers a celebrity has, a post made by he/she influences a large population. While others argue that many other factors come into picture when quantifying how influential is a person rather than raw follower count. This can be analyzed by collecting a large volume of Twitter data and identifying the key factors that contribute to the influence of a user. Many methods are in use for collecting data from Twitter with the most popular ones being Twitter API and GOT3 or using readily

available data online. Identifying influenced users across Twitter can be helpful in multiple domains. For example, this analysis can be used to formulate marketing strategies by organizations so that they can approach the more influenced people to make a post on their products for them to reach out to a large population. Also identifying most influenced persons during election campaigns can be used to predict the winning presidential candidate. Candidates can even deploy strategies to make themselves more influential on Twitter which indeed helps in their campaigning.

That data for the analysis is obtained from Kaggle data science competition. The train datasets has 5500 rows and 13 columns. The approach here is to classify each of users which has attributes from both the A user and the B user into the classes making it a Binary classification problem. The data is split into train and test datasets using 5 fold cross-validation. This is because we do not have test data and to test the model performance we need test data to work with. Many approaches for binary classification such as Multilayer Neural Network, Logistic regression, Random Forest Classifier, Support Vector Classifier(SVC), Stochastic Gradient Descent(SGD), K Nearest Neighbours(KNN), Naive Bayes have been used to classify every row in the data into either of the two classes. The models are fitted on the train part of the split dataset and tested on the testing data set for obtaining the model accuracy.

Prior to fitting the models, it is important to obtain the features that are most useful for classification. Using all the features in the dataset is not a good approach as the model tends to memorizes the data points which eventually leads to overfitting. Random forest feature importance is used to rank the features in the dataset based on their importance towards the classification task. While there are many approaches towards feature reduction, Random forest is most followed approach giving the best possible results for the data in many situations. Then a subset of these features of the variable importance is taken for model fitting and discarding the remaining features. The random forest approach is run multiple amounts of times before feature selection since there is randomness involved in the approach and it is optimal to normalize the results from various iterations.

The above-stated models are then fitted on this subset of features in an iterative fashion. Various parameters of these models must be taken into consideration when fitting the models. The parameters should then be tweaked according to the data set and obtain the best parameter set for every model. Once the models are fitted it is important to evaluate the performance of these models and compare them against one another to obtain the best model that fits the data. The performance is tested on the test data obtained from the split using metrics such as accuracy score, precision score, recall score, F1 score and confusion matrix. These are various metrics that describe various properties of the performance of the model.

## 2 LITERATURE REVIEW

Previous research on determining the user influence in twitter by Krishna P. Gummadi suggests that the indegree, reweets, and mentions play a major role in determining a user's influence on Twitter. Where, in-degree refers to how popular a user is, re-tweets is the number of re-tweets received by a post made by the user and mentions is the number of mentions he got from his post. Krishna P. Gummadi says that a user being more popular may not be equally influential in terms of re-tweets and mentions. In his paper, he only used these three attributes to determine the influence of a user on Twitter.[5]

Whereas, the approach proposed currently uses more attributes than that and hence is expected to perform better than Krishna P. Gummadi's model. It is always better to have more attributes or more data so that we can then perform feature selection to select the most important features. We can also calculate the correlations between different features to see how they influence each other. It is better to consider re-tweets or mentions received and sent and also the follower as well as the following count to determine the influence to model a better classifier.[5]

Katz, E., and Lazarsfeld discuss a useful method to determine the most influential customer using social network so that the companies can market their product to so that they can market it to an exponential number of people. They say that if instead of viewing a market as a set of independent entities, we view it as a social network and model it as a Markov random field. Their method focuses on calculating a network for a given customer. The current method proposed is an extension to this, predicting who is the most influential user, instead of calculating the network value of each user.[7]

## 3 DATA DESCRIPTION

The data for classification and analysis is taken from Kaggle's competition on "Influencers In Twitter". The idea was to use bigger data set, but due to scalability issues, have only used a smaller data set. The data set has 5500 rows and 23 columns. Each row in the data corresponds to two users A and B. Each row is independent of one another i.e. the A user and B user in each row is independent of each other. The first column 'Choice' in the data set represents the class label. This column has a value of 1 if B user is more influential than the A user and has a value of 0 if A user is more influential than the B user. The next 11 columns belong to attributes of the A user followed by 11 columns belonging to the attributes of B user. The following are the 11 attributes of each user:

- (1) Follower-count
- (2) Following-count
- (3) Listed-count
- (4) Mentions-received
- (5) Mentions-sent
- (6) Retweets-received
- (7) Retweets-sent
- (8) Posts
- (9) Network feature-1
- (10) Network feature-2

- (11) Network feature-3

**Follower count, Following count:** These attributes in the data specifies the number of users following the A or B user and number of users followed by user A or B respectively.

**Listed count:** This attribute specifies the number of private lists the A or B user is a part of.

**Mentions Received and Retweets Received:** These attributes specify the number of mentions and the number of re-tweets A or B user received from other users.

**Re-tweet Sent:** specifies the number of re-tweets A or B user sent to other users.

**Post:** This attribute describes the number of tweets made till date by the users.

**Network feature attributes:** These are obtained as a result of PCA analysis and are included in the data. They correspond to some variance in the data from the network point of view.

## 4 FEATURE SELECTION

Due to the existence of a large number of features available in the data set, this feature should be reduced and only the most important features that help to model the data needs to be selected. If all the features are used to model the data the model performs exceptionally well on the training data set but fails to perform well on the testing data set. This is called 'Over-fitting'. This is a general phenomenon that occurs across most Machine learning algorithms when using a large set of features for classification because the model learns to memorize these features resulting in a good accuracy score for train data set, but when we encounter a new set of values for these features in the testing data the models fail to classify them correctly resulting in a poor accuracy score.

There are many approaches to feature selection. One approach is to have domain knowledge on the data i.e. having prior knowledge on the data set and by intuition selecting features from the data. For example, if we are predicting sales prices of a house based on the features of the house, having an in-depth knowledge on what feature make the house price go up or down can help in selecting the features of interest. This often requires a professional who experts in that domain to make the decision. But this approach often fails for two reasons, firstly it is hard to find an expert opinion on every data set and secondly the intuition of the expert can be trusted as humans are prone to make mistakes in estimation.

The ideal approach would be to use some algorithm that can predict the features of interest by iterating over the model multiple times. Random forest for feature selection is one such algorithm. As the word random suggests, the algorithm builds small decision trees using a different subset of features in each iteration and then combines the insights from all of these decision trees and ranks the features in the dataset based on their importance. Since the algorithm builds thousands of trees based on the random selection of features, the feature ranking, for the most part, is accurate. Once the features with its ranking are obtained, one has to select a subset of these features and use them for model classification. Since the ideal number of features that best suits the model is still unknown, the approach is to iterate over the important feature and select the best possible subset. For the Twitter dataset, the random forest ranked the features according to their importance and selecting 8 of

the top most important feature given the best possible results. The features are follower count, listed count, re-tweets received and Network features 1 for both A and B users. Figure 1 shows the features ranked according to their importance.

[Figure 1 about here.]

## 5 CLASSIFICATION METHODS

Since we have only two classes (either 0 or 1) in our class label, it becomes a binary classification problem. A lot models can be used for binary classification. The approach is to fit 8 of these models for the data set and evaluate the performance of each of these models by using the evaluation metrics.

### 5.1 Logistic Regression

Logistic regression is one of the Generalized Linear Models that describes the relationship between the response variable and the explanatory variables[1]. The response variable can either be binary or multinomial. If the response variable is binary, it is called binary logistic regression and if it is multinomial, it is called multinomial logistic regression.

The cumulative distribution function for logistic regression is given by:

[1]

$$F(x) = P(X \leq x) = \frac{e^{\frac{(x-\mu)}{\tau}}}{1 + e^{\frac{(x-\mu)}{\tau}}} \quad (1)$$

Where  $\mu$  is the mean of the given set of data points and  $\tau$  is a scaling parameter.

Using the cumulative distribution function for logistic regression[1], the logistic regression function is given by:

[1]

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \quad (2)$$

where:

[1]

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \quad (3)$$

$\pi$  is the probability of occurrence of an event. Where

[1]

$$\alpha = -\mu/\tau \quad (4)$$

$$\beta = 1/\tau \quad (5)$$

The relationship between  $\pi(x)$  and  $x$  is mostly non-linear. The influence of  $x$  on  $\pi(x)$  is more when it is near 0 or 1 rather than when it is in between.  $\pi$  either increases or decreases with an increase in  $x$ [1]. The rate of increase or decrease depends on  $\beta$ . If  $\beta > 0$ , then  $\pi(x)$  increases with an increase in  $x$  and if  $\beta < 0$ , then  $\pi(x)$  decreases with an increase in  $x$  and remains constant when  $\beta = 0$ .

The random component for the response variable in logistic regression has a binomial distribution. The link function is 'logit'. The logistic regression model is also sometimes referred to as logit model. If it is a multinomial logistic regression model, then the random component would be multinomially distributed[1].

### 5.2 Random Forest Classifier

The Random forest algorithm apart from its use as feature selection method can also be used a Classifier. Random forest in one of the ensemble classification methods currently in use."Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem"[11]. Typical methods for example Decision tree algorithm build trees by iterating over the features in the train data set. But ensemble method uses a different approach, for example, Random forest build a number of small decision trees based on a subset of training features and combine the results from all these decision trees and obtain a better accuracy. It eventually builds a forest of decision trees, hence the name Random Forest.[2]

Each of the decision trees has a decision as to which class a new data point belongs to. Random forest algorithm classifies the data point by combining the decisions from all the decision trees. Hence it provides better performance while at the same time over-fitting on most occasions. The random forest classifier has two stages, the training phase, and the classification phase. [2]

In the training phase from the corpus of training features, a random subset of features is sampled with replacement. Three-fourths of the data is sampled leaving out the remaining data for formulating the classification error. It is called out of bag data (OOB). In the classification phase, proximities are calculated for every decision tree for each of the two classes. If two decision trees have the same proximities for each of the two classes the proximity is increased by one. Once all the proximities are determined it is then divided by the total number of decision trees to normalize it.[2]

While building the decision trees typically the Gini importance is used.Gini index splits the tree based on whether the Gini impurity criterion of the current node is less than the parent node.

[2]

$$I_G(p) = 1 - \sum_{i=1}^J (p_i)^2 \quad (6)$$

Where  $I_G(p)$  is gini impurity of p

J is the number of classes (J=2 for binary classification)  $p_i$  is the number of items classified into class i

Here there are a few design choices to be made

**Max iterations:** - Have to specify the maximum number of iterations the classifier has to run. It is hard to find an optimal number for every problem. Need's to be iterated across multiple values.

**Max depth:-** Specifies the maximum depth to which the decision trees are built.

**N estimators:-** specifies the maximum number of trees built in each iteration.

**Max features-** specifies the number of features to be considered when splitting the decision tree.

**Criterion** - specifies using either gini impurity index or entropy. By default it uses gini impurity index.[2]

### 5.3 Stochastic Gradient Descent

Gradient descent is one of the machine learning algorithms which is used for updating the weights of a neural network to optimize the prediction error of machine learning algorithms[4]. The weights

are updated so that the classification or regression error of training samples is minimized over the iterations. Gradient descent is classified into three types based on the number of training samples used for updating the weights. They are:

- (1) Batch Gradient Descent
- (2) Stochastic Gradient Descent
- (3) Mini-batch Gradient Descent

In batch Gradient descent, the weights of the network are assigned randomly initially and the model is designed by updating the weights accordingly to minimize the prediction error after all the training samples are classified. [4]

Stochastic Gradient Descent is another type of gradient descent in which the error is calculated for each training sample. The weights are updated for every training sample. Because of the kind of update the Stochastic Gradient Descent uses, it is called as online machine learning algorithm. There are both advantages and disadvantages of using this approach. One of the advantages is that the model is less prone to arrive at local minima as the update process is noisy[4]. As the error is calculated frequently, the error in the prediction can be corrected at that stage itself, instead of propagating it to next stages. Also, the frequent updating of weights may result in fast learning. On the other hand, one of the major disadvantages of this approach is that updating the weights for every training sample may be much time consuming and delay the convergence. Also the updates being noisy might cause the algorithm to arrive at a error minimum with high variance[4].

To combine the advantages of both Batch Gradient Descent and Stochastic Gradient Descent, Mini-batch Gradient Descent can be used[4]. In this method, the training data is divided into a number of subsets and the error is calculated and weights are updated for each subset.

#### 5.4 Support vector Classifier

Support Vector Machines is an advancement over Neural Network. A neural network converges if the two classes are linearly separable. Every time we run a neural network on the train data and plot the points and the line on a 2-D plane, we get a different line that separates the two classes, this happens due to the randomness in the initial weights chosen for the perceptrons. Since we get a different classifier every time we run it, it is obvious that one classifier is better than the other. The neural network does not guarantee the best possible classification. This is where SVC shines with its improvements over neural networks. SVC can be used both for classification and regression tasks but are typically used for classification. There mainly three types of SVC,[10]

**1.Linear SVC**

**2.Nu SVC**

**3.SVC**

Ideally, the goal of SVC is to place the decision boundary i.e. the line that separates the two classes as far away from the classes as possible. The distance between the decision boundary and the data points in each class is called margin. Most often some of the data points exist within the margin and are close to the boundary, these points are called as 'Support vectors'. The distance from the vector to the boundary is given by,[10]

$$r = \frac{g(x)}{\|w\|} \quad (7)$$

where,

$$g(x) = w^T x + b \quad (8)$$

where r is the distance, g(x) is the decision boundary and  $\|w\|$  is the absolute value of the weight vector and b is the bias.

Since the goal here is to find the optimal decision boundary, the corresponding w and b should be estimated. This can be achieved by using the Primal problem which is a constraint minimization problem. The formulation of this is,[10]

$$d_i(w^T x_i + b) \geq 1 \quad (9)$$

$$\Phi(w) = \frac{1}{2} w w^T \quad (10)$$

The above minimization can be solved using the Lagrangian formulation,

$$J(w, b, \alpha) = \frac{1}{2} w w^T - \sum_{i=1}^N \alpha_i [d_i(w^T x_i + b) - 1] \quad (11)$$

where  $\alpha_i$  is called the Lagrangian multipliers. We obtain the optimal values of w and b by partially deriving the above equation and equating it to zero. We get,

$$w = \sum_i \alpha_i d_i x_i \quad (12)$$

$$Q(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j d_i d_j x_i^T x_j \quad (13)$$

here  $\alpha_i > 0$  only for support vectors and Q is quadratic. Finally after getting  $\alpha_i$ , the optimal  $w_0$  is,

$$w_0 = \sum_{i=1}^{N_s} \alpha_0 d_i x_i \quad (14)$$

where  $N_s$  is the number of support vectors.

$$b_0 = \frac{1}{d(s)} - w_0^T x^{(s)} \quad (15)$$

Since now we have the optimal value for w and b, we get the optimal decision boundary. SVC generally gives a better accuracy over perceptron and it can be used efficiently for high dimensional planes.[10]

#### 5.5 Multi layer Perceptron

The neural network came into existence in the early 1960's but was not very popular approach since it had some limitations in solving complex problems. But with the introduction of backpropagation algorithms and its application on Multi-layer Neural networks. Each neural network comprises of individual neurons which are also called perceptions. These neurons are connected to one another with edges. A simple neural network has three layers, input layer,

an output layer and the hidden layer. The inputs layers consist of data points these data points are connected to the neuron in the hidden layer by an edge which has weights and biases associated with them. The hidden layer is then connected to the output layer with edges which also has weights and biases.[9]

Based on the type of the problem, the output layer either output a binary value(0 or 1) or any continuous value. The input data point is multiplied by the weight of the neuron and the bias of that neuron is added. All output from each neuron is added together and is fed to an activation function and produces an output. Once an iteration is completed the weights of the network are updated based on the output obtained from the first iteration using the perceptron learning algorithm. The learning stops when there is no significant improvement from one iteration to the other. The error is calculated after each iteration which quantifies the number of misclassified data points.[9]

$$w(n+1) = w(n) + \eta[d(n) - y(n)] * x(n) \quad (16)$$

where,

w(n+1)-new weight of the neuron  
w(n)- old weight of the neuron  
d(n)- desired output  
y(n)- acutal output  
x(n)- data point

This simple neural network can only be used yo solve a handful of problems. For example, it cannot solve the XOR problem. To deal with much more complicated problems we require the use of Multi-layer Neural networks with the back-propagation algorithm to update the weights in each iteration. The back-propagation algorithm uses gradient descent to update the weights by cost minimization. There are two steps in the back-propagation algorithm. One is the feedforward step where the network feeds forward and find the output. In the back-propagation step the algorithm back-propagates and update the weights in each layer by using gradient descent. This is done is a backward fashion starting from the output layer and then the hidden layers. There are two methods for updating the weights, batch update and the online update. While batch update method updates the weights after one iteration of all the data points, the online update method updates the wights after each data point is passed through the network.[9]

Here there are lot of design choice to made

- 1.Initializing weights - weights are usually initialized random normal manner.
- 2.Choosing the number of hidden layers- This parameter cannot be determined in advance. One has to iterate through multiple values of the hidden layer unless the best accuracy is obtained.
- 3.Update method- The weights can be updated typically using the batch update on the online update. In most application typically the online method is used as it is much more efficient.
- 4.Choosing the Activation function- This is a difficult task, most commonly non-linear activation functions are used. Example Re-Lu activation function. Figure 2 shows the architecture of a Multi-layer Neural network

[Figure 2 about here.]

## 5.6 K Nearest Neighbours

K Nearest Neighbours is one of the simplest machine learning models that can be easily implemented and it surprisingly works well for most classification problems. K Nearest Neighbours can be used for both classification and regression problems giving it the flexibility to solve most problems.

The algorithm assumes all the data points are scattered on a 2-dimensional plane and it tries to find clusters of these data points. The algorithm mainly has two phases, the training phase, and the classification phase. The training phase of KNN is fairly simple as it just stores the values of each of the variables and also the class label. The classification phase is the most important one as all the calculation are made in this phase. Since cross-validation is being used to split the data set into train and test, in the training phase the algorithm stores the values of each of the 8 attributes used for model fitting and also the values of the class label for each fold of the cross-validation.

In the classification phase for each data point in the test data set, the algorithm iterates over all the data point in the train data set and finds the k nearest neighbors to the current data point by calculating the distance between them. A variety of distance metrics can be used to find the nearest neighbors. The most commonly used ones are Manhattan Distance and Euclidean Distance. But based on the type of data other distance metrics can be used to obtain better performance. Once the k nearest neighbors are obtained the algorithm looks for the class labels for these neighbors and assigns the most occurring label among these neighbors to the current data point. This happens for every data point in the test data set and every fold in the cross-validation. The algorithm then assigns the most occurring class labels across all folds.

The value of k is unknown for a given problem and one has to iterate over multiple values of k to find the best accurate model.The most common approach to find k is to iterate over multiple values of k starting from 1 and stopping when the difference in performance from the previous value of k is below a certain threshold. This threshold can be set based on the type and volume of data. Ideally, the threshold should not be too large or too small. Some advantages of this model involve low training cost and it often works well with enough training data and a good distance metric. It also has few drawbacks which involve choosing the apt distance metric can be cumbersome without prior knowledge, it is easily prone to over-fitting and usually takes a lot of time and memory for the classification phase.

## 5.7 Naive Bayes

Nave Bayes is one of the machine learning algorithms. It is known for its simplicity. It is a classification approach based on the Bayefis law. In this approach, prior knowledge is very essential. The basic idea is to compute the prior of each class and the likelihoods. Then, the posterior probabilities are calculated using Bayefis law for each class and the data point is assigned to the class with highest posterior probability. The prior represents the probability of occurrence of a class among all the given classes. The likelihoods or likelihood probabilities are a measure of how likely is a given data point prone to belong to a given class. Posterior probability is the probability

of a class is one of the given classes, given a set of data points.[6] There are two phases in the Nave Bayefis classification as that of any other machine learning algorithm, training, and testing. In the training phase, the prior probabilities of the classes are calculated by computing the number of times the class has occurred in the given data among all the classes. If there are n classes  $C_1, C_2, \dots, C_n$  and m data points  $X_1, X_2, \dots, X_m$ , then the prior of a class is represented by  $P(C_i)$  for i ranging from 1 to n. After computing the priors, the likelihood of each data point belonging to each one of the classes are calculated. The likelihood ratio is given by:[6]

$$P(X_j/C_i) = \frac{P(X_j, C_i)}{P(C_i)} \quad (17)$$

where  $i = 1$  to  $n$ ,  $j = 1$  to  $m$

Where,  $P(X_j, C_i)$  is the probability of occurrence of the data point  $X_j$  in class  $C_i$ .  $X_j$  represents the  $j^{th}$  data point in a given set of data points and  $C_i$  represents the  $i^{th}$  class.[6]

In the testing phase, the posteriors probabilities are using the Bayefis law:

$$P(C_i/X_j) = \frac{P(X_j/C_i)P(C_i)}{P(X_j)} \quad (18)$$

Where,  $P(X_j)$  represents the total probability of  $X_j$ , given by:

$$P(X_j) = \sum_{i=1}^n P(X_j/C_i)P(C_i) \quad (19)$$

Naive Bayes model makes the assumption that the data points are conditionally independent of each other given the class or label. Hence, the likelihood probability of a set of data points belonging to a class becomes:[6]

$$P(X_1, X_2, \dots, X_n/C_i) = \prod_{j=1}^n P(X_j/C_i) \quad (20)$$

Then the posterior probability of a set of data points can be written as:

$$P(C_i/X_1, X_2, \dots, X_n) = \frac{\prod_{j=1}^n P(X_j/C_i)P(C_i)}{P(X_1, X_2, \dots, X_n)} \quad (21)$$

For a given set of data points,  $P(X_1, X_2, \dots, X_n)$  remains constant and hence, equation(6) can be written as:

$$P(C_i/X_1, X_2, \dots, X_n) \propto \prod_{j=1}^n P(X_j/C_i)P(C_i) \quad (22)$$

Naive Bayes derives its name from making the naive assumption that the data points are conditionally independent of each other. There are three types of Naive Bayes classifiers:[6]

- (1) Gaussian Naive Bayes
- (2) Multinomial Naive Bayes
- (3) Bernoulli Naive Bayes

Gaussian Naive Bayes can be used for the classification when the given set of data points have Gaussian distribution.[8]

Multinomial Naive Bayes can be used when for the data if it is multinomially distributed[8]. Bernoulli Naive Bayes can be used when all the attributes or features in a given data are binary[8].

## 5.8 AdaBoost

Ensemble learning is one of the machine learning methods which is based on decision trees. The basic idea is to create an ensemble of hypotheses and combine their predictions instead of predicting using a single hypothesis[3]. Boosting is one of the ensemble methods which attempts at learning a strong classifier from a set of weak classifiers. At first, an initial model is created and then the next model attempts to correctly classify the training samples which have been misclassified by the previous model. This process continues until a predefined number of models have been generated or there is nothing much to be done with the training data.

AdaBoost is one of the Boosting algorithms which is used for binary classification. It can only be used when the response variable is binary. AdaBoost can be used with any machine learning algorithms to improve the performance. It is most commonly used with decision trees. The decision trees used in AdaBoost are only of depth one, i.e., they only have one decision to make. Hence, the decision trees in AdaBoost are called decision stumps[3].

There are two phases in the AdaBoost as well, like any other machine learning algorithms. In the training phase, all the given training samples are assigned with uniform weights. If there are n training samples, then each sample is given a weight of  $1/n$ . Hence, the initial weights of the training samples can be written as[3]:

$$w(x_i) = \frac{1}{n} \quad (23)$$

Where  $x_i$  is the  $i^{th}$  training sample and  $w(x_i)$  is the initial weight of  $i^{th}$  training sample. Using these weighted samples, a weak binary classifier is modeled which can only make one decision and output either -1 or 1, where 1 represents the first class and -1 represents the second class. Then, the misclassification rate is calculated for the trained model as follows[3]:

$$\text{error} = \frac{(c - n)}{n} \quad (24)$$

Where c is the number of correctly classified samples and n is the total number of training samples. Now, weighted aggregate of the misclassification rate is calculated to further use it to modify the weights of the training samples. It is modified as[3]:

$$\text{error} = \frac{\sum_{i=1}^n w(x_i)e(i)}{\sum_{i=1}^n w(x_i)} \quad (25)$$

Where  $e(i)$  is the prediction error of the  $i^{th}$  training sample which is either 1 or 0. It is 0 if the sample is classified correctly and 1 if it is misclassified.

Then, stage value is calculated from the aggregated error as follows[3]:

$$\text{stage} = \ln\left(\frac{1 - \text{error}}{\text{error}}\right) \quad (26)$$

The stage value is used to assign greater importance to classifiers which are more accurate. The weights of the training samples are updated using the stage value so that samples which are correctly classified have less weight and the incorrectly classified samples have more weight. The weights are updated using the below equation[3]:

$$w(i) = w(i) * e^{\text{stage}*e(i)} \quad (27)$$

We can observe from the above equation that if a sample i is correctly classified, then prediction error of that sample  $e(i)$  will be 0

and hence the weight of the sample remains as it is. Whereas, if the sample is misclassified, then  $e(i)$  would be 1 and the weight of the sample increases by a factor  $e^{stag}$ . This process is continued by updating the weights and modeling weak classifiers until a predefined number of weak classifiers have been modeled or when there is nothing more to be gained from the training data.[3]

In the testing or prediction phase, when a test sample is to be classified into one of the two available classes, it is classified using all the weak classifiers designed in the training phase and the predicted values are given weights according to the stage value of the corresponding classifiers[3]. The final predicted value is taken to be 1 if the weighted sum of these predicted values is positive and -1 if the sum is negative.

## 6 EXPLORATORY DATA ANALYSIS

### 6.1 Missing values

Since the data is sparse, null values can exist in the data. These null values if not treated can be catastrophic to the model accuracy. The null or missing values can be treated either by removing the rows which have these values or replacing them with apt data. For example, if a data frame has 10 null values in a column, the approach would be to replace the values with either the mean, median or mode of that column. The data here does not consist of any null or missing values in any of the rows.

### 6.2 Outliers

If a data point diverges very much from the rest of the data points it is called an outlier. These outliers if not addressed will eventually lead to overfitting of the model i.e. the model fits each and every point in the train data and gets a good accuracy, but when tested on unseen data it will result in a poor accuracy score. To overcome overfitting the outliers need to be addressed. The outliers in the data can be identified by either plotting a box plot or scatter plot of each of the variables in the train data and check for the distribution of data points.

Just like missing values the outlier can either be deleted from the data or can be modified to better suit the model. The ideal approach would be to use binning of the columns which have outliers. In binning data, we bin the data into small categories based on a range. The Twitter data here consists of a few outliers in almost all the attributes. Binning the data improved the accuracy score of all the models. Figure 3 shows the scatter plot of A follower count variable with outliers.

[Figure 3 about here.]

From the scatter plot it is evident that for all the attributes the data point is concentrated in the lower range. Hence after binning each attribute into two the first two lower categories, null values are generated in the higher categories. Hence the rows with Null values have been dropped thereby removing the outliers. The data is reduced to 5209 rows after the removal of outliers.

### 6.3 Correlation among attributes

Since the data consists of 22 columns apart from the class label, correlations among these attributes might be. Correlation is the

measure of association between two attributes. It indicates how closely two attributes are related to each other. The range of correlation is between -1 and 1. A positive correlation indicates the two variables are directly proportional and a negative correlation indicated that they are indirectly proportional. If the value is zero then there is no association between the attributes. Figure 4 shows a heat map displaying the correlation among the variables.

[Figure 4 about here.]

The color of the heat map indicates the level of correlation among attributes. As the color tends to get closer to yellow, it shows a positive correlation and shows a negative correlation if color tends to get bluer. For example, there is a strong correlation between network feature 1 and mentions received. This correlation also helps in feature selection as we can ignore one of the two features if they are closely related. The variables which are closely related to the class label are found to be follower count and listed count of both A and B. This is not very surprising as a person with high follower count or listed count on Twitter is expected to be more influential than others. These insights proved to be useful in the further analysis when performing feature reduction.

## 7 EXPERIMENT AND RESULTS

To avoid overfitting, the training data has been further split into train and test using cross-validation. 80 percent of the data has been used for training the classifier and the remaining 20 percent is used to evaluate the performance of the classifier.

Feature selection has been performed on the train data using Random Forest Classifier to obtain the most important attributes. A bar plot has been plotted to know the top most important variables which are shown in the figure. The top four important attributes are identified to be follower-count, listed-count, mentions-received and network feature-1 of both A and B. Further analysis or classification has been performed on these eight attributes.

### 7.1 Evaluation Metrics

Once the model is fitted, there is a need to evaluate the performance of the model. Various evaluation metric can be used to measure how well the model performs. Accuracy score, precision score, F1 score, Confusion matrix and recall score are the most commonly used metrics to evaluate the performance of classifiers. They indicate various measure of evaluation

1. **Accuracy score:** Ratio of number of correctly classified data points to total number of data points.
2. **Precision score:** Ratio of True positives to sum of True positives ,False positives for each class.
3. **F1 score:** Twice the ratio of product of precision and recall to the sum of precision and recall for each class.
4. **Recall score:** Ratio of True positives to sum of True positives ,False negatives for each class
5. **Confusion matrix:** To know how well the classifier performs in each class in the data. It gives the number of True positive, False positives, True negatives, and False negatives. The rows in the matrix correspond to actual class label and the columns to the predicted class label.

If a classifier has very high precision then it may miss many true instances of a label, similarly, if it has very high recall score it is prone to misclassify many data points. Hence a good classifier should have a trade-off between precision score and recall score. For the Influence analysis, the Ada boost Classifier has the best performance with an accuracy score of 0.78 and an F1 score of 0.78. The MLP classifier performed the worst with an accuracy score of 0.68 and an F1 score of 0.67. This is because the classifier performed extremely well in one class and very poorly in the other class. This is evident from a high precision score in class 1 and a low recall score, and vice-versa for class 0. Whereas for the AdaBoost classifier both class 1 and 0 have decent precision and recall scores, hence finding a good trade-off between them. Figure 5 shows the evaluation performance of each classifier with all the evaluation metrics

[Figure 5 about here.]

## 8 FUTURE WORK

Predicting the Influence of users based on various attributes paves the way for much more research in the area. Instead just classifying two users based on their influence it can be extended to a larger audience, which can be quite interesting. But it becomes a regression problem rather than a classification problem. This analysis can also be extended to address how an individual is influential i.e. in a good or a bad way. This can be achieved with the help of sentiment analysis performed on the users. But this would require the actual tweets that the users make.

## 9 CONCLUSION

Twitter being one of the most popular social media platforms, serves as a source for various thoughts expressed by people across different domains. These tweets or tweet data if analyzed in an efficient way can answer many questions. One of such questions is which user on Twitter is more influential. Given the tweets made by the user and other statistics like a number of retweets received, follower count, etc the user with more influence can be predicted using various machine learning classification techniques. AdaBoost is found to be the most efficient classification technique for this research question with an accuracy score of 0.78 and the least appropriate method was found to be the Multi-Layer Perceptron classification.

## ACKNOWLEDGMENTS

We would like to thank Dr. Gregor von Laszewski and the AIs for all the help they have provided for this project.

## A WORK BREAKDOWN

**Dataset identification:** Bharat Mallala

**Exploratory data analysis:** Bharat Mallala and Jyothi Pranavi Devineni

**Feature Selection:** Jyothi Pranavi Devineni

**Random Forest, SVM, K nearest neighbours, MLP :** Bharat Mallala

**Naive bayes, SGD, Logistic regression, Adaboost:** Jyothi Pranavi Devineni

**Project Report:** Bharat Mallala and Jyothi Pranavi Devineni

## REFERENCES

- [1] Alan Agresti. 2007. *Introduction to Categorical Data Analysis*. Vol. 2. JohnWiley & Sons, Inc., Hoboken, New Jersey. 91–94 pages.
- [2] Leo Breiman and Adele Cutler. 2010. Rnandom Forests. (2010). [https://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm)
- [3] Jason Brownlee. 2016. Boosting and AdaBoost for Machine Learning. (2016). <https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/>
- [4] Jason Brownlee. 2017. A Gentle Introduction to Mini-Batch Gradient Descent. (2017). <https://machinelearningmastery.com/gentle-introduction-mini-batch-gradient-descent-configure-batch-size/>
- [5] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi (Eds.). 2010. *Measuring User Influence in Twitter: The Million Follower Fallacy*. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1538/1826>, 2011
- [6] David Crandall. 2017. Lecture on Naive bayes. (2017). [https://iu.instructure.com/courses/1649672/files/72107879?module.item\\_id=16147477](https://iu.instructure.com/courses/1649672/files/72107879?module.item_id=16147477)
- [7] E Katz and Lazarfeld. 1955. Personal Influence: The Part Played by People in the Flow of Mass Communications. (1955).
- [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [9] Donald Williamson. 2016. Lecture on Neural Networks. (2016). <https://iu.instructure.com/courses/1600135/files/folder/Lecture%20Slides?preview=69588803>
- [10] Donald Williamson. 2016. Lecture on Support Vector Machines. (2016). <https://iu.instructure.com/courses/1600135/files/folder/Lecture%20Slides?preview=70070725>
- [11] Zhi-Hua Zhou. 2016. Ensemble Learning. *National Key Laboratory for Novel Software Technology* 23, 122-127 (2016), 220–235.

#### LIST OF FIGURES

1	Variable importance	10
2	Architecture of a Multi-layer Neural network [9]	11
3	Scatter Plot	12
4	Correlation Plot	13
5	Evaluation Metrics	13

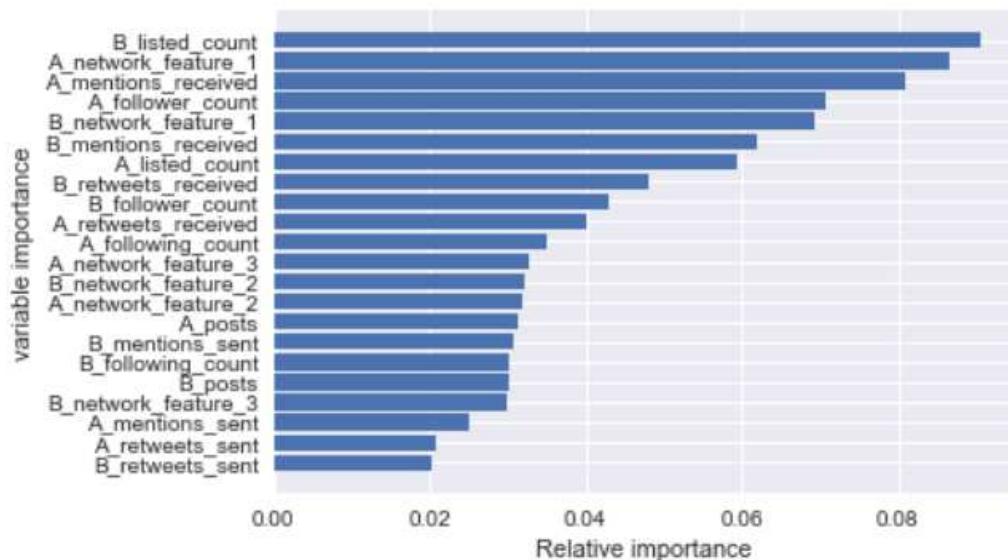


Figure 1: Variable importance

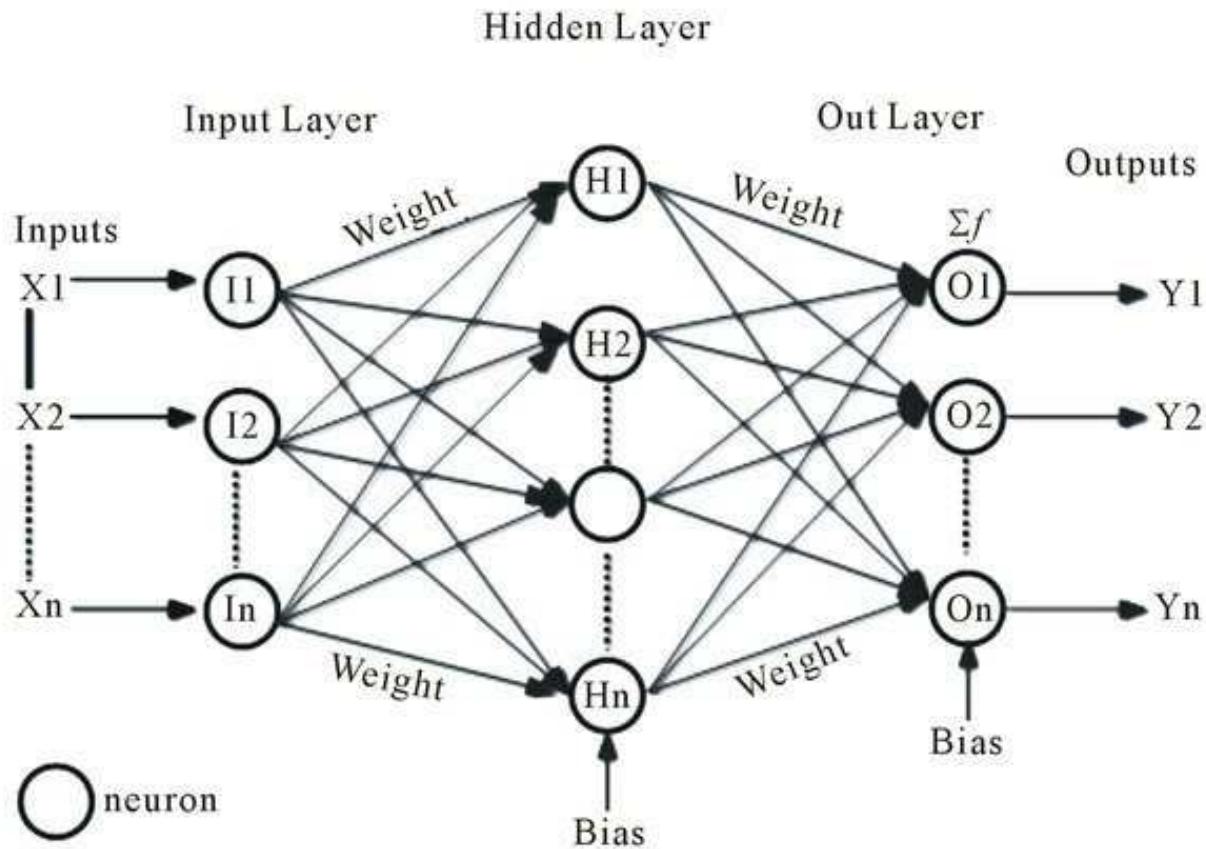


Figure 2: Architecture of a Multi-layer Neural network [9]

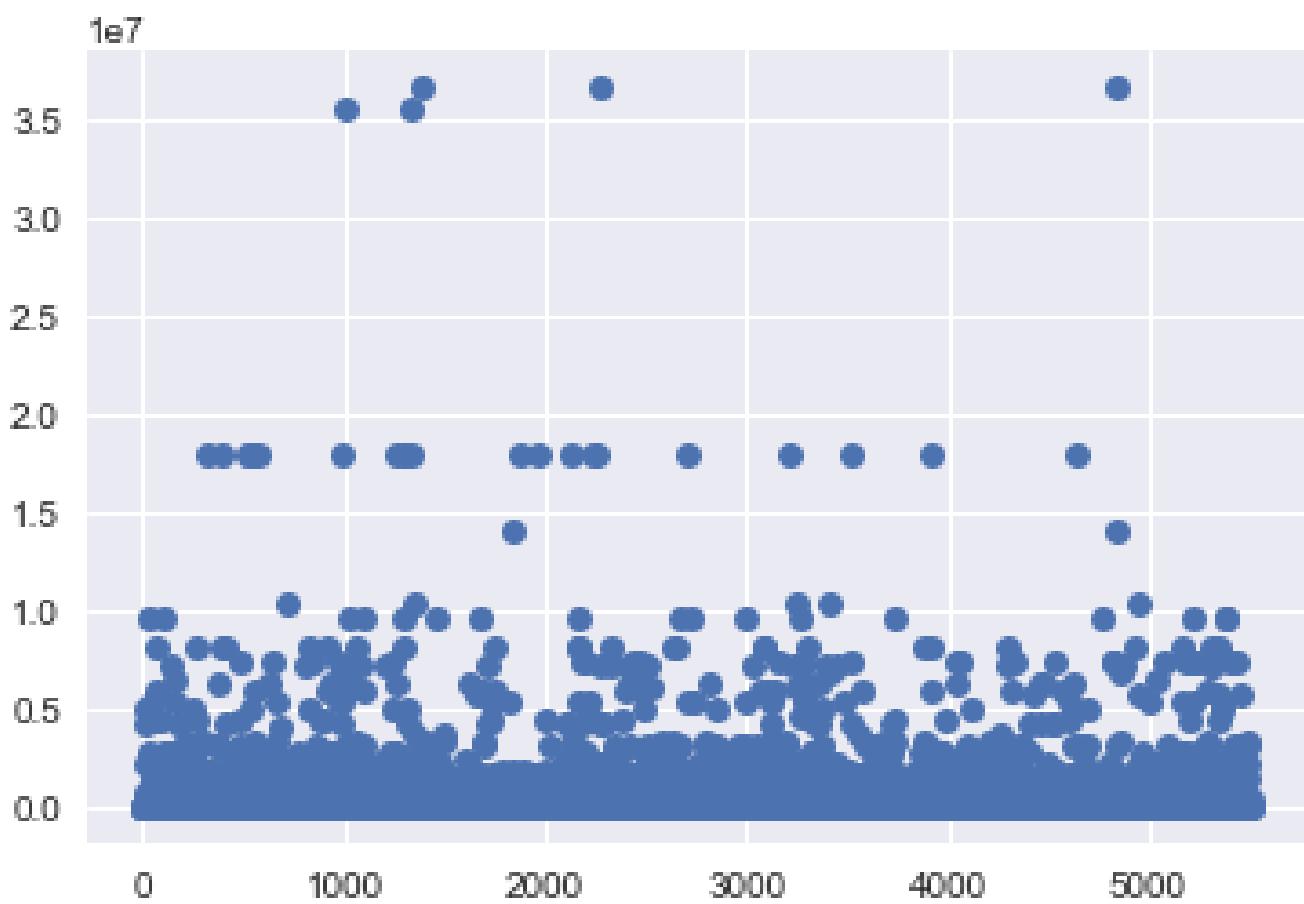


Figure 3: Scatter Plot

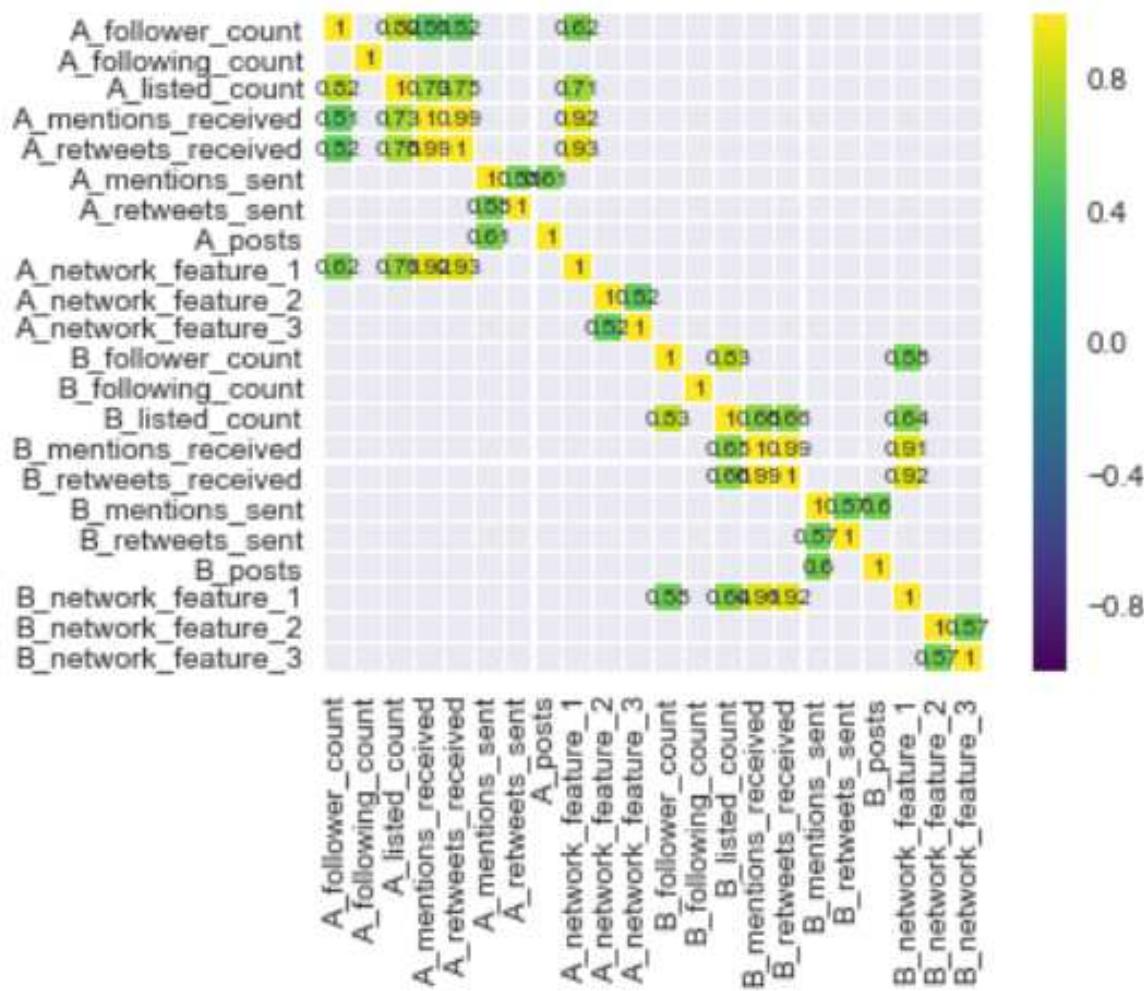


Figure 4: Correlation Plot

	accuracy score	precision score	recall score	f1 score	confusion matrix
<b>logistic regression</b>	0.774472	[0.76334519573, 0.7875]	[0.80790960452, 0.739726027397]	0.774143	[[429, 102], [133, 378]]
<b>Knn</b>	0.744722	[0.747663551402, 0.741617357002]	[0.75329566855, 0.735812133072]	0.744699	[[400, 131], [135, 376]]
<b>Random forests</b>	0.776392	[0.762323943662, 0.793248945148]	[0.815442561205, 0.735812133072]	0.775956	[[433, 98], [135, 376]]
<b>LinearSVC</b>	0.772553	[0.761565836299, 0.785416666667]	[0.806026365348, 0.737769080235]	0.772221	[[428, 103], [134, 377]]
<b>Adboost</b>	0.786948	[0.804733727811, 0.770093457944]	[0.768361581921, 0.80626223092]	0.786929	[[408, 123], [99, 412]]
<b>Multinomial NB</b>	0.757198	[0.756457564576, 0.758]	[0.772128060264, 0.74168297456]	0.757121	[[410, 121], [132, 379]]
<b>SGD</b>	0.759117	[0.748226950355, 0.771966527197]	[0.79472693032, 0.722113502935]	0.758728	[[422, 109], [142, 369]]
<b>MLP</b>	0.687140	[0.92531120332, 0.615480649189]	[0.419962335217, 0.964774951076]	0.662954	[[223, 308], [18, 493]]

Figure 5: Evaluation Metrics

## bibtex report

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

bibtext \_ label error

bibtext space label error

bibtext comma label error

latex report

[2017-12-16 09.32.53] pdflatex report.tex

```
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.3s.
./README.yml
 7:14     warning  truthy value is not quoted  (truthy)
```

---

## Compliance Report

---

```
name: Mallala, Bharat
hid: 215
paper1: Oct 29 17 100%
paper2: Nov 6 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 215 project (null) 7249 report.tex
wc 215 project (null) 7221 report.pdf
wc 215 project (null) 354 report.bib
```

```
find "
```

---

64: The data for classification and analysis is taken from Kaggle's competition on "Influencers In Twitter". The idea was to use bigger data set, but due to scalability issues, have only used a smaller data set. The data set has 5500 rows and 23 columns. Each row in the data corresponds to two users A and B. Each row is independent of one another i.e. the A user and B user in each row is independent of each other. The first column 'Choice' in the data set represents the class label. This column has a value of 1 if B user is more influential than the A user and has a value of 0

if A user is more influential than the B user. The next 11 columns belong to attributes of the A user followed by 11 columns belonging to the attributes of B user.\\"

133: The Random forest algorithm apart from its use as feature selection method can also be used a Classifier. Random forest is one of the ensemble classification methods currently in use."Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem"\cite{Zhou2016}. Typical methods for example Decision tree algorithm build trees by iterating over the features in the train data set. But ensemble method uses a different approach, for example, Random forest build a number of small decision trees based on a subset of training features and combine the results from all these decision trees and obtain a better accuracy. It eventually builds a forest of decision trees, hence the name Random Forest.\cite{Breiman2010}\\\

passed: False

find footnote

---

16: \renewcommand\footnotetextcopyrightpermission[1]{} % removes footnote with conference information in first column

passed: False

find input{format/i523}

---

6: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

92: The features are follower count, listed count, re-tweets received and Network features 1 for both A and B users. Figure \ref{fig:Fig1} shows the features ranked according to their

```

importance. \\

94: \begin{figure}
95: \includegraphics[width=1.0\columnwidth]{images/fig1.png}
97: \label{fig:Fig1}

220: 4.Choosing the Activation function- This is a difficult task,
      most commonly non-linear activation functions are used. Example
      Re-Lu activation function. Figure \ref{fig:Fig2} shows the
      architecture of a Multi-layer Neural network \\

222: \begin{figure}
223: \includegraphics[width=1.0\columnwidth]{images/fig2.jpg}
225: \label{fig:Fig2}

301: Just like missing values the outlier can either be deleted from
      the data or can be modified to better suit the model. The ideal
      approach would be to use binning of the columns which have
      outliers. In binning data, we bin the data into small categories
      based on a range. The Twitter data here consists of a few
      outliers in almost all the attributes. Binning the data improved
      the accuracy score of all the models. Figure \ref{fig:Fig3} shows
      the scatter plot of A follower count variable with outliers. \\

303: \begin{figure}
304: \includegraphics[width=1.0\columnwidth]{images/fig3.png}
306: \label{fig:Fig3}

312: Since the data consists of 22 columns apart from the class label,
      correlations among these attributes might be. Correlation is the
      measure of association between two attributes. It indicates how
      closely two attributes are related to each other. The range of
      correlation is between -1 and 1. A positive correlation indicates
      the two variables are directly proportional and a negative
      correlation indicated that they are indirectly proportional. If
      the value is zero then there is no association between the
      attributes. Figure \ref{fig:Fig4} shows a heat map displaying the
      correlation among the variables. \\

314: \begin{figure}
315: \includegraphics[width=1.0\columnwidth]{images/fig4.png}
317: \label{fig:Fig4}

332: For the Influence analysis, the Ada boost Classifier has the best
      performance with an accuracy score of 0.78 and an F1 score of
      0.78. The MLP classifier performed the worst with an accuracy
      score of 0.68 and an F1 score of 0.67. This is because the
      classifier performed extremely well in one class and very poorly
      in the other class. This is evident from a high precision score
      in class 1 and a low recall score, and vice-versa for class 0.
      Whereas for the AdaBoost classifier both class 1 and 0 have
      decent precision and recall scores, hence finding a good trade-
      off between them. Figure \ref{fig:Fig5} shows the evaluation
      performance of each classifier with all the evaluation metrics\\

```

```
334: \begin{figure}
335: \includegraphics[width=1.0\columnwidth]{images/fig5.jpeg}
337: \label{fig:Fig5}
```

```
figures 5
tables 0
includegraphics 5
labels 5
refs 5
floats 5
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

---

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8217
non ascii found 8217
non ascii found 8220
non ascii found 8221
non ascii found 239
non ascii found 8217
non ascii found 8217
non ascii found 239
non ascii found 8217
non ascii found 8230
non ascii found 8230
non ascii found 8217
```

---

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# New Approaches to Managing Metadata at Scale in Research Libraries

Timothy A. Thompson

Indiana University Bloomington

School of Informatics, Computing, and Engineering

Bloomington, Indiana 47408

timathom@indiana.edu

## ABSTRACT

The analysis of big data often relies on distributed storage and computation; however, access to big data—and to the platforms capable of managing and processing it—continues to be largely centralized. Centralization is particularly evident in the case of the metadata produced, managed, and disseminated by academic and research libraries. Libraries typically create and share their catalog records by uploading them to a central data warehouse, which can then be searched by other libraries for records that can be copied and added to an institution’s local catalog. Centralization has the advantage of scalability and availability, but it comes at the cost of a loss of autonomy. Existing metadata workflows can be optimized through the adoption of entity resolution and machine learning algorithms, including approaches based on neural network models. Although innovation is possible within the current paradigm, it can only reach its full potential in the context of peer-to-peer platforms that would allow libraries to share their data directly.

## KEYWORDS

i523, HID340, Research Libraries, Library Catalogs, Entity Resolution, Neural Networks

## 1 INTRODUCTION

The problem of entity resolution (also known as record linkage or data matching [? ]) is one that has a direct impact on the work of information professionals in research libraries. In library units responsible for catalog management, many workflows center on a procedure known as copy cataloging, which aims to expedite the processing of new acquisitions. Copy cataloging involves searching a shared database for records created by another cataloging agency, but that describe identical publications that have been acquired locally [? ]. In the current environment, a single company, the Online Computer Library Center (OCLC—<http://www.oclc.org>), is the only viable platform for global cooperative cataloging [? ]. OCLC provides data aggregation and warehousing services that allow libraries to effectively share their data, but its business model does not encourage peer-to-peer interaction and innovation among individual libraries. This centralized model, which operates on the basis of membership fees, has the advantage of scalability and availability, but it comes at the cost of a loss of control over the data itself, and it entails the acceptance of a business model that, in effect, charges libraries for serving their own data back to them.

## 2 NEW APPROACHES TO METADATA MANAGEMENT

Libraries have a tradition of experience with record matching and automation [? ], but now stand to benefit from the increasingly mainstream availability of algorithms and routines developed within the context of data science and machine learning. Sophisticated algorithms for string comparison and probabilistic data record linkage have long been available, but are not widely used by libraries, with the exception of large-scale projects such as the Social Networks and Archival Context Project (SNAC) (<http://snaccooperative.org/>) and the Virtual International Authority File (VIAF) (<http://viaf.org/>). The former has employed methods based on Naïve Bayes classification algorithms to aggregate and disambiguate data from across a wide range of libraries and archives. The reported accuracy of this approach fell with the range of 80-90 percent.

### 2.1 Neural Networks for Data Matching

## 3 CONCLUSION

The blockchain-based database BigchainDB (written in Python) provides an alternative, and to benefit from features such as data immutability and an asset-based transactional model. A working prototype installation of a BigchainDB node has the potential to provide an example of how libraries can abandon centralized models for managing their data at scale.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the i523 teaching assistants for their support and suggestions in writing this report.

## REFERENCES

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "pC12"
Warning--I didn't find a database entry for "cD17"
Warning--I didn't find a database entry for "aT10"
Warning--I didn't find a database entry for "jM92"
(There were 4 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-16 09.39.00] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.1 L31 : [pC12] undefined
p.1 L31 : [cD17] undefined
p.1 L31 : [aT10] undefined
p.1 L34 : [jM92] undefined
Empty 'thebibliography' environment.
There were undefined citations.
Typesetting of "report.tex" completed in 0.8s.
```

```
=====
Compliance Report
=====
```

```
name: Tim Thompson
hid: 340
```

```
paper1: Oct 25 17 100%
paper2: 100% Nov 8 17
project: Dec 7 17 66%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
1
wc 340 project 1 607 report.tex
wc 340 project 1 573 report.pdf
wc 340 project 1 730 report.bib
```

```
find "
```

---

35: Libraries have a tradition of experience with record matching and automation \cite{jM92}, but now stand to benefit from the increasingly mainstream availability of algorithms and routines developed within the context of data science and machine learning. Sophisticated algorithms for string comparison and probabilistic data record linkage have long been available, but are not widely used by libraries, with the exception of large-scale projects such as the Social Networks and Archival Context Project (SNAC) (\url{http://snaccooperative.org/}) and the Virtual International Authority File (VIAF) (\url{http://viaf.org/}). The former has employed methods based on Naive Bayes classification algorithms to aggregate and disambiguate data from across a wide range of libraries and archives. The reported accuracy of this approach fell with the range of 80-90 percent.

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
7: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
figures 0
tables 0
includegraphics 0
labels 0
refs 0
floats 0
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "pC12"
Warning--I didn't find a database entry for "cD17"
Warning--I didn't find a database entry for "aT10"
Warning--I didn't find a database entry for "jM92"
(There were 4 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# How Far Have Spacewalks Walked

Rick Alan Carmickle Jr.

Indiana University School of Informatics and Computing

919 E. 10th Street

Bloomington, IN 47408, USA

rcarmick@iu.edu

## ABSTRACT

We use data from the NASA Open Data Portal mission parameters parsed from the Wikipedia pages of manned spaceflight missions of the U.S. and Russian space programs to analyze American and Russian spacewalks. We calculate the distance traveled, relative to the earth's surface, by astronauts and cosmonauts who performed Extravehicular Activities (EVA) in earth orbit. As part of that calculation, numerous other orbital parameters will be calculated.

## KEYWORDS

i523, HID304, Big Data, Wikipedia API, Extra Vehicular Activity

## 1 INTRODUCTION

This was an exercise in using datasets from unrelated sources, cleaning that data so that links between datasets could be formed, then creating mathematical analysis and visualizations of the data from necessary elements in these different datasets. The data used was a combination between an existing dataset created by NASA as part of the NASA Open Data initiative, and mission trajectory datasets created by gathering and cleaning data on manned space launches via the Wikipedia API.

The goal was to measure the distances traveled through space by humans in earth-orbit spacewalk. At the outset, an objective of this project was to create a geovisualization of an orbital ground trace for the specific area spacewalks occurred. The author discovered that several pieces of data necessary to perform the objective were missing from Wikipedia data. This paper will end with a discussion of exactly what data is needed, in addition to what is available here, to create orbital ground traces and highlight portions of those traces.

## 2 EXPLANATION OF TERMS

Terms related to orbital mechanics and the data sources need to be defined at the outset.

Spacewalk and Extravehicular Activity are synonymous. This refers to events when a human in a vessel outside earth's atmosphere leaves the interior or opens the hatch of a vessel exposing the interior to space [1].

The Infobox of a Wikipedia page is the table on many pages with facts and summaries of the article's topic. This is an embedded object on Wikipedia pages with standard syntax but many variations on how it can be used [13].

The Apogee or Apoapsis of an orbit is the orbit's maximum distance from the body it is orbiting in an elliptical path. Apogee refers specifically to objects in Earth's orbit and Apoapsis is a general term. Perigee or Periapsis of an orbit is the orbit's minimum distance from the body it is orbiting in an elliptical path [15]. Perigee

refers specifically to objects in Earth's orbit and Periapsis is the general term. Both of these measurements are typically recorded as distances from the surface of the orbited body [8].

Orbital eccentricity is a measurement of the difference between the apoapsis and periapsis. Put another way, it is a measurement of how elliptical an orbit is. An eccentricity of 0.0 indicates a perfectly circular orbit. An elliptical orbit will have an eccentricity of between 0.0 and 1.0. Eccentricity greater than or equal to 1.0 indicates an escape trajectory [8].

Orbital inclination is a measurement of the angle between an object's orbital plane and the earth's equatorial plane. An inclination of 0.0 is a perfectly equatorial orbit traveling east relative to the Earth's surface: the same direction as the Earth's rotation. An inclination of 90.0 is a polar orbit and an inclination of 180.0 is a perfectly equatorial orbit traveling west against the Earth's rotation [8].

## 3 GETTING THE DATA

The starting point is the dataset from the NASA Open Data Portal, a 2015 dataset entitled "Extra-vehicular Activity (EVA) - US and Russiafifi [9]. This dataset contains plaintext of the country, crew, vehicle, duration, and purpose of each spacewalk. The official start of each space walk, defined as the moment an astronaut or cosmonaut has begun exiting an open hatch in a depressurized vessel, is included as a date.

From the fiVehiclefi category of this data, a single column of unique mission names was generated manually. This single column was formatted to remove non-mission-related text and appended to a single-category csv file. The 'Vehicle' category formatting was vital to this project since this was the merging category for the EVA dataset and orbital parameter dataset.

The Jupyter notebook '1-gathering-wikipedia-infobox-data' contains the steps for creating a dataset with significant details on each given mission which can be generated from the a single category containing the name of the mission.

The data gathering stage of the project emerged, over the course of this project, as the step with the highest Big Data potential. The usability and flexibility of the code developed here was unexpected for the author and has demonstrated applicability for gathering and parsing Wikipedia data for other desired categories.

Several different open-source libraries were used together at this stage in a notebook capable of identifying and sorting categories of data from JSON templates contained on given Wikipedia pages [10]. This notebook is specific to the categories of interest in this project because the names of the Infobox parameters desired must be specifically given as arguments.

These parameters could, however, could be altered to fit any set of Infobox categories. The interface for searching and retrieving

JSON data on Wikipedia pages was tested during development with wikipedia pages unrelated to manned space travel and is capable of retrieving data from pages with different JSON configurations on data charts contained on Wikipedia.

The fission-namefi file is the only necessary input file to generate the dataset. The only requirement of the category names in this file is that they must be close to the string that would appear as the title of the desired Wikipedia page. They do not, however, need to be an exact match. The 'Wikidiapi library by Jonathan Goldsmith [5] is a flexible library which can access the Wikipedia API and utilize Wikidiapis search function to retrieve the URL of the desired page. This library had many functions for accessing Wikipedia data, but does not include a method to access tables and charts.

This library was capable of identifying than an input of "Gemini IVfifi corresponds to the proper article title of "Gemini 4fifi and the exact URL string of "<https://en.wikipedia.org/wiki/Gemini-4fifi>. The exact article title string is necessary for the step which iterates through individual pages and appends the data.

The data fetching process makes a request through pywikibot [12] of the Wikipedia API for a given page. Pywikibot retrieves a raw version of the entire page. Mwparserfromhell [7] identifies templates matching JSON charts and returns a name and value list for every parameter name and values for every JSON chart template on each page.

The input file which provides article titles and the category names matched to the parameter names in the JSON date could be changed to retrieve data for other interests. The code uses for gathering the data is not efficient and had much room for improvement. A future project the author will pursue is an interface which accepts a column of strings corresponding to entities with Wikipedia pages and a column of parameter names from the Infobox and returning a dataset which fills as much data as is available.

## 4 CLEANING THE DATA

The Jupyter notebook '2-data-cleaning-and-merging' cleans both the EVA and raw mission datasets. The fission-data-raw.csvfi data is not a usable dataset immediately following parsing from Wikipedia. The author was unable to implement code to recognize raw JavaScript and JSON data to append only the necessary substrings, so this textual data was cleaned with code written specifically for each category. This stage relies heavily on basic string editing functions such as replace, strip, and split. For maximum flexibility with calculations, elements of the data were reduced to the simplest components or summarized in a way that left a single integer or floating value. For example, the original string scraped from Wikipedia for fission durationfi was roughly formatted as fin months, x, days, y hours, z minutesfi but permutations of JavaScript code, misread ASCII characters, and inconsistencies in the categories used were common. Several categories of data were added in the data cleaning process. One instance of data creation which should be noted is the appending of current International Space Station (ISS) orbital parameters to all missions beginning with fiExpeditionfi. This data should be considered an experimental stand-in for complete data for respective spacewalks because the ISSfi orbit is not stable. The orbit of the ISS decays over time since

the craft experiences a small amount of atmospheric drag [6]. The ISS is also navigated in orbit. The true orbital parameters of the ISS for the duration of any given spacewalk would require a data framework which can access the orbital history of the ISS, a function beyond the scope of this project and the skillset of the author. The ISS has, however, remained close to 250 miles overhead for the duration of its existence and has never been accelerated to create a high orbital eccentricity, so a generalized set of orbital parameters will provide an acceptable baseline for ISS spacewalks.

The data cleaning for many categories was a repetitive task of removing errant characters. Future development would benefit from the author better understanding the relationship between JavaScript, JSON, html, and Python so that syntax from one language can be cleaned before importing to a data file.

The mission data parsed from Wikipedia was not the only dataset that needed addressing. The EVA dataset required work to render it usable. This dataset included many one-time typos and oddities in columns which required manual changing in the csv data. fiExtra-vehicular-Activity-EVA-US-and-Russia.csvfi is therefore not exactly as one would find it from the NASA Open Data Portal. The precise alterations made to this data are detailed in fichangelog to Extra-vehicular-Activity-EVA-US-and-Russia.txtfi. The altered EVA datafile is needed to run this project.

New categories of data were generated from the EVA dataset before it was ever merged with the orbital parameter data. The single string of crew names was split into separate columns, each containing either the name of a crew member who participated in EVA or fiNaNfi. The spacewalk duration, in seconds, was calculated for ease in mathematics later on. Over the course of this project, the author learned much about orbital ground tracks and was refreshed on the mathematics underlying orbital mechanics. One of the limitations of the EVA dataset was revealed at this stage: The dataset includes the data (month/day/year) where a given spacewalk occurred, but the creation of an orbital EVA ground trace would require the month, day, year, hour, minute, and second of the beginning of an EVA to create an accurate trace given the relative speed over the ground of low-earth orbit spacecraft. The available data does provide information accurate enough to calculate orbital velocity and distances traveled.

The data cleaning notebook ends with merging the EVA data and the orbital parameter data into a single DataFrame on the fiVehiclefi column.

## 5 DATA ANALYSIS AND CALCULATIONS

The third notebook uses only the new dataset finspacewalk-with-orbital-data.csvfi. Because the data has been cleaned, merged, and homogenized to consistent units of measurement, the mathematics of further orbital details are relatively simple calculations across a DataFrame. The Earthfis equatorial radius and gravitational constant [14] are variables contained within the notebook and units are converted as needed. From the data scraped from the Wikipedia API, we can calculate the semi-major and semi-minor axes of each orbit, the eccentricity, and the orbital velocity of a given orbits where all earlier parameters were present [15].

## 6 RESULTS

The initial question of "How far have spacewalks walked" is answered with the data generated here, with the caveat that input data is imperfect. The longest single spacewalk, in terms of time spent in EVA and distance traveled relative to the earth, is a 2001 Space Shuttle Mission working on the ISS where Jim Voss and Susan Helms worked outside for 8 hours and 56 minutes and traveled 246,946,540 meters relative to the earth's surface. Table ?? includes the descriptive EVA data as well as the final calculated fields for the twenty most traveled spacewalks.

Figure 1 features the final calculated spacewalk distances for every spacewalk. Only the mission label is included. Repeats of a label indicate multiple EVAs by a mission's crew. The distances traveled range from the roughly 5,500 kilometers traversed by Alexei Leonov over 12 minutes in the very first human spacewalk to the quarter of a million kilometers traveled by Mission Specialists Jim Voss and Susan Helms in their 8 hour and 56 minute spacewalk as part of an early resupply mission to the ISS.

Figure 2 is a parallel coordinate plot of the year a given EVA took place mapped to EVA duration. From this visualization, we see the increasing volume of spacewalks as well as the general increase in EVA time over the years as the technology to sustain low-earth orbit EVAs improved and space program efforts were focused on the construction of the ISS.

Upon basic exploration of the data, we see some trends by mission type. The most eccentric orbits were often the earliest, with Gemini, Voskhod, and Apollo appearing as the most eccentric. STS and Gemini missions appear to be the missions with the largest orbital perimeters. The highest orbital velocities, directly related to how close to the surface an object orbits, are topped by Soyuz missions from the late 1980s through the early 2000s as well as STS missions.

A glance at the orbital inclination reveals a distinct pattern in this category. Data points are clumped around the values of 28 and 50. Figure 3 is a scatterplot of spacewalk duration and orbit inclination, colored by country [11]. There is a reason for the distinct clumping in the data along those values.

The inclination of an orbit is a measurement of the object's orbital plane compared to earth's equatorial plane. An orbiting object's inclination can never start at less than the latitude of the launch site. Lowering the inclination of an orbit below the starting latitude requires an adjustment burn at either orbital insertion or during orbit when latitude matches the desired inclination [3]. In Figure 3, the lower line of datapoints, at a value around 28 corresponds to the Latitude of Kennedy Space Center, which is the source of most American launches and is located at 28°28'N 80°32'W. The upper clumping of data points, at a value around 52 corresponds to the ISS inclination of 51.64 degrees.

The predicted and actual orbital period categories present an opportunity to test the validity of the calculations made with this dataset. We would expect an ideal 1-to-1 relationship between predicted and actual orbital periods. The predicted and actual periods are, in this instance, independent variables since predicted period was calculated with Apoapsis, Periapsis, semimajor axis, and constant values of earth's properties [8, 15]. The actual orbit

period measurement was not part of the calculation. The difference between these values has a mean of 6.265 seconds, a median of 0.210 seconds, and a standard deviation of 38.75. With single columns of mathematically independent variables, a linear regression is appropriate [2]. Figure 4 displays the results of the linear regression model. The p-value of this model is 1.824 u-150. This is a demonstration of an effective relationship between variables as expected. In this model, the slope and intercept would carry some significance. The ideal relationship between two identical variables should have an intercept of 0 and a slope of 1 [4]. The intercept of 441.38 is 7.996 percent of the mean of actual orbital periods. The slope of 0.9211 is 7.89 percent less than the expected slope. Both slope and intercept [4] are almost precisely 8 percent deviated from the ideal expectation R-value and standard error confirm that the calculated orbital period is deviated by approximately 8 percent from the true values reported in the data [4].

The geovisualization aspect of this project ultimately failed. After beginning the process of creating orbital ground traces through BaseMap, it was discovered that a precise ground trace requires some additional data either unavailable or unparsed from the Wikipedia data. A ground tracing function would require the time, to the second or minute, of both the beginning of the spacewalk and of the mission. The parameter fields recognized the original data parsing code reuse the field name "Date" for the multiple dates in the JSON tables on each mission page, including the date the page was published, which is irrelevant to this project. Further understanding of the JSON and html is needed before we could complete this step. Two other pieces of data needed for accuracy are the earth coordinates of the point over which the vessel first established an orbit with apoapsis and the decay rate for orbits which still experienced atmospheric drag.

The analysis notebook includes code to begin the process of creating correct ground traces. The data parsed from Wikipedia lacked several categories needed to complete this aspect of the exercise. The data fields needed were launch date and times, EVA date and times, and decimal coordinates for launch and landing sites. This data is available, but not through the data source used in this exercise. Additional data scraping methods are needed to complete that exercise, but the Wikipedia data alone enabled a close effort.

## 7 FIGURES

In Figure 1 we show a complete list of distance traveled in EVA by astronauts of every American or Russian mission launched until 2013.

[Figure 1 about here.]

In Figure 2 we show a parallel coordinate plot of mission launch year and EVA duration.

[Figure 2 about here.]

In Figure 3 we show a scatterplot of mission inclination colored by country of launch. A pattern reflective of launchsites and mission destinations is apparent.

[Figure 3 about here.]

In Figure 4 we show the linear regression results of the reported orbital period versus the predicted period calculated from other fields in the data.

[Figure 4 about here.]

## 8 TABLES

[Table 1 about here.]

## 9 CONCLUSION

The distances traveled by spacewalkers are closely related, but not directly calculated from, the durations of spacewalks. This calculation required many fields before it could be determined. The inability to ultimately construct a correct ground trace visualization was disappointing, but a great exercise in both python's Basemap library and a for a more accurate understanding of orbital mechanics. The Wikipedia data parsing by category presents an interface through which other data projects may be pursued and something the author will return to in future work.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and his course TAs for their support.

## REFERENCES

- [1] National Aeronautics and Space Administration. 1995. *Extravehicular Activity (EVA)* (revision b ed.). National Aeronautics and Space Administration. <https://msis.jsc.nasa.gov/sections/section14.htm>
- [2] The Scipy Community. 2017. `scipy.stats.linregress`. [docs.scipy.org](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html). (Oct. 2017). <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html>
- [3] Jason Davis. 2012. Of inclinations and azimuths. The Planetary Society. (April 2012). <http://www.planetary.org/blogs/guest-blogs/jason-davis/3450.html>
- [4] Jim Frost. 2014. How to Interpret a Regression Model with Low R-squared and Low P values. The Minitab Blog. (June 2014). <http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values>
- [5] Jonathan Goldsmith. 2017. python wikipedia API. <https://github.com/goldsmith/Wikipedia>. (2017). <https://pypi.python.org/pypi/wikipedia>
- [6] Lee Hutchinson. 2013. How NASA steers the International Space Station around space junk. arstechnica. (July 2013). <https://arstechnica.com/science/2013/07/how-nasa-steers-the-international-space-station-around-space-junk/>
- [7] Ben Kurtovic. 2017. mwparserfromhell 0.5. <https://github.com/earwig/mwparserfromhell/tarball/v0.5>. (2017). <https://pypi.python.org/pypi/mwparserfromhell>
- [8] James Muirden. 1982. *Astronomy Handbook*. Arco.
- [9] NASA. 2014. Extra-vehicular Activity (EVA) - US and Russia. NASA's Open Data Portal. (July 2014). <https://data.nasa.gov/Raw-Data/Extra-vehicular-Activity-EVA-US-and-Russia/9kcy-zwvn>
- [10] stackoverflow.com/questions. 2012. Content of infobox of Wikipedia. stackoverflow.com/questions. (June 2012). <https://stackoverflow.com/questions/8088226/content-of-infobox-of-wikipedia>
- [11] Michael Waskom. 2017. seaborn.FacetGrid. <http://seaborn.pydata.org/generated/seaborn.FacetGrid.html>
- [12] Wikimedia. 2017. Pywikibot. <https://github.com/wikimedia/pywikibot>. (2017). <https://www.mediawiki.org/wiki/Manual:Pywikibot/Installation>
- [13] Wikipedia. 17. `Template:Infobox`. (Feb. 17). [https://en.wikipedia.org/wiki/Template:Infobox#Infoboxes\\_and\\_user\\_style](https://en.wikipedia.org/wiki/Template:Infobox#Infoboxes_and_user_style)
- [14] Dr. David R. Williams. 2016. Earth Fact Sheet. nasa.gov. (Dec. 2016). <https://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html> NASA Official: Dr. David R. Williams, [david.r.williams@nasa.gov](mailto:david.r.williams@nasa.gov).
- [15] Richard Wolfson. 2007. *Essential University Physics* (first edition ed.). Vol. 1. Pearson Addison Wesley.

Included in the Appendix are the three Jupyter notebooks used in this project.

## A JUPYTER NOTEBOOKS

you need to include citations and describe what the notebooks do

### A.1 Jupyter Notebook 1

`1-gathering-wikipedia-infobox-data.ipynb`

### A.2 Jupyter Notebook 2

`2-data-cleaning-and-merging.ipynb`

### A.3 Jupyter Notebook 3

`3-spacewalk-analysis.ipynb`

This is not how we introduce float figures, see your other figures

`/includegraphics[width=\columnwidth]{images/myimage.pdf}`

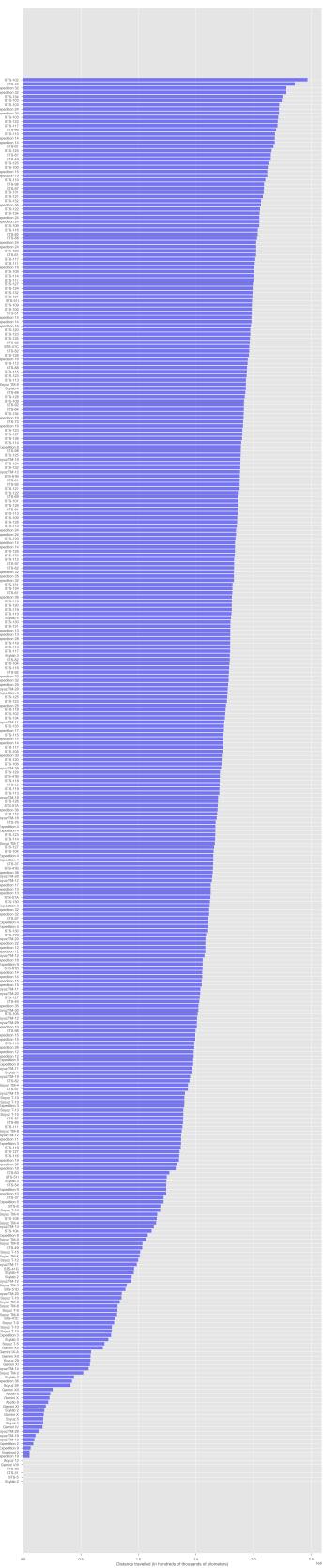
paper is real short, insufficient explanation provided to some tables and figures.

there should not be a section called TABLES, FIGURES, you need a result section.

LIST OF FIGURES

1	All Distances	7
2	Parallel Coordinate Launch Year and EVA	8
3	Inclination and Country	9
4	Linear Regression	10





7  
**Figure 1: All Distances**

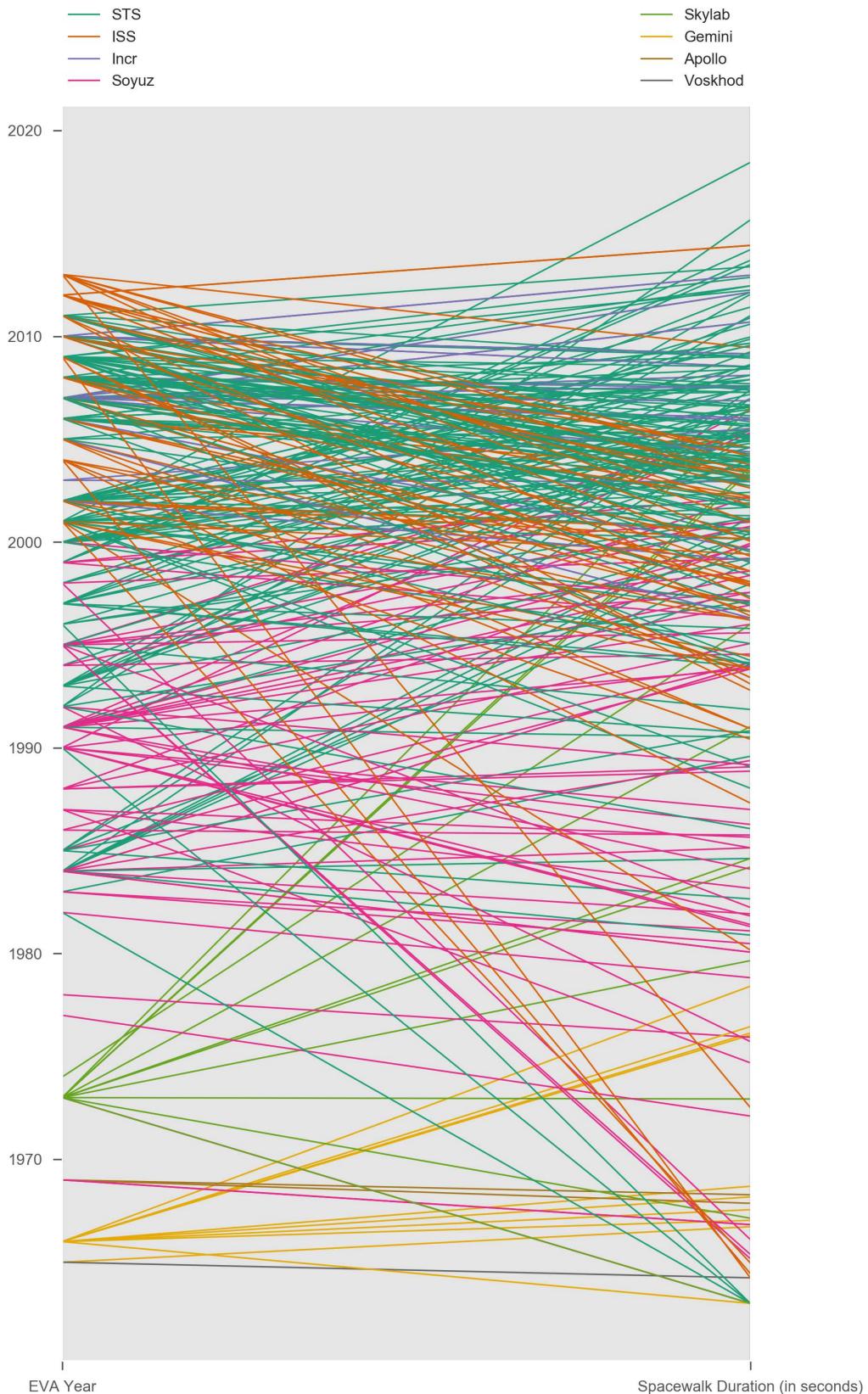
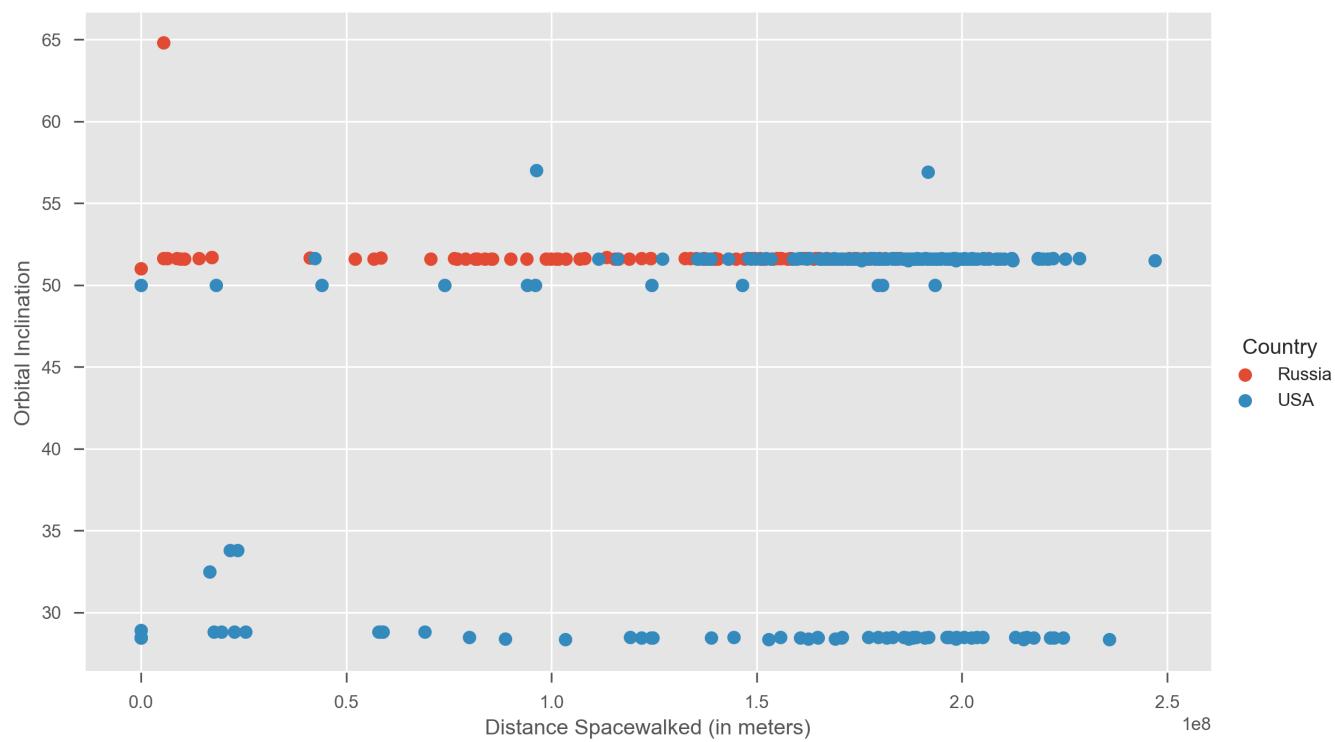
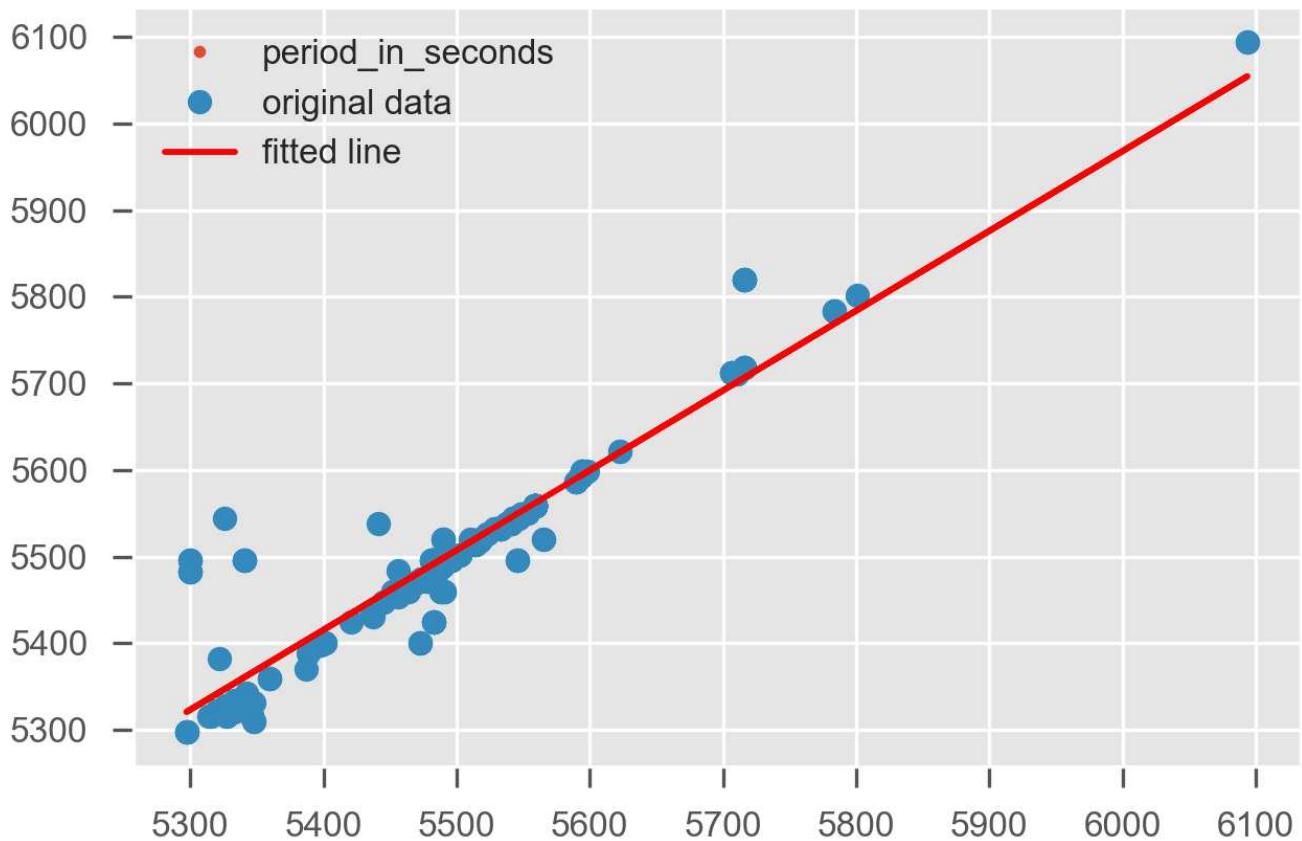


Figure 2: Parallel Coordinate Launch Year and EVA<sup>8</sup>



**Figure 3: Inclination and Country**



**Figure 4: Linear Regression**

LIST OF TABLES

1 Table 1. Longest Distance Spacewalks

12

**Table 1: Table 1. Longest Distance Spacewalks**

Vehicle	Date	Duration	Distance Spacewalked (meters)	Crew 1	Crew 2	Crew 3	Spacewalk Duration (seconds)	Orbital Perimeter	Orbital Velocity
STS-102	3/10/2001	8:56	246946540	Jim Voss	Susan Helms	32160	42432.42	7678.69	
STS-49	5/13/1992	8:29	235881646	Pierre J. Thuot	Richard J. Hieb	Thomas D. Akers	41986.01	7723.70	
Expedition 32	8/30/2012	8:17	228596876	Sunita Williams	Akihiko Hoshide	29820	42614.69	7665.89	
STS-134	5/22/2011	8:07	225199861	Andrew Feustel	Mike Fincke	29220	42159.08	7707.05	
STS-103	12/22/1999	8:15	224675143	Steve Smith	John Grunsfeld	29700	43754.92	7564.82	
STS-103	12/23/1999	8:10	222405697	Mike Foale	Claude Nicollier	29400	43754.92	7564.82	
Expedition 24	8/7/2010	8:03	222157527	Doug Wheelock	Tracy Caldwell Dyson	28980	42614.69	7665.89	
STS-103	12/24/1999	8:08	221497919	Steve Smith	John Grunsfeld	29280	43754.92	7564.82	
STS-122	2/11/2008	7:58	220991575	Rex Walheim	Stan Love	28680	42177.95	7705.42	
STS-117	6/15/2007	7:58	220876174	Danny Olivas	Jim Reilly	28680	42234.48	7701.40	
STS-96	5/29/1999	7:55	219611347	Tammy Jernigan	Dan Barry	28500	42165.38	7705.66	
STS-110	4/11/2002	7:48	218717662	Steve Smith	Rex Walheim	28080	41266.60	7789.09	
Expedition 14	1/31/2007	7:55	218477900	Mike Lopez-Alegria	Sunita Williams	28500	42614.69	7665.89	
STS-61	12/4/1993	7:54	217400720	Story Musgrave	Jeff Hoffman	28440	42792.17	7644.19	
STS-125	5/17/2009	8:02	215733575	Mike Massimino	Mike Good	28920	43415.26	7459.67	
STS-87	11/24/1997	7:43	215075057	Winston Scott	Takao Doi	27780	41807.25	7742.08	
STS-49	5/14/1992	7:44	215027670	Tom Akers	Kathy Thornton	27840	41986.01	7723.70	
STS-125	5/15/2009	7:56	213048095	Mike Massimino	Mike Good	28560	43415.26	7459.67	
STS-100	4/24/2001	7:40	212401275	Scott Parazynski	Chris Hadfield	27600	42290.94	7695.70	
Expedition 15	7/23/2007	7:41	212038551	Clay Anderson	Fyodor Yurchikhin	27660	42614.69	7665.89	

## bibtex report

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--can't use both author and editor fields in Muirden1982
Warning--empty address in Muirden1982
Warning--can't use both author and editor fields in Wolfson2007
Warning--empty address in Wolfson2007
(There were 4 warnings)
```

## bibtext \_ label error

bibtext space label error

report.bib:70: note = {NASA Official: Dr. David R. Williams, [david.r.williams@nasa.gov](mailto:david.r.williams@nasa.gov)}

## bibtext comma label error



```
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Float too large for page by 23.0pt.
Float too large for page by 23.0pt.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
There were undefined references.
Typesetting of "report.tex" completed in 6.5s.
./README.yml
 8:47      error    trailing spaces (trailing-spaces)
 9:50      error    trailing spaces (trailing-spaces)
 10:48     error    trailing spaces (trailing-spaces)
 11:49     error    trailing spaces (trailing-spaces)
 12:48     error    trailing spaces (trailing-spaces)
 13:47     error    trailing spaces (trailing-spaces)
 14:47     error    trailing spaces (trailing-spaces)
 15:49     error    trailing spaces (trailing-spaces)
 16:48     error    trailing spaces (trailing-spaces)
 17:49     error    trailing spaces (trailing-spaces)
 18:51     error    trailing spaces (trailing-spaces)
 19:45     error    trailing spaces (trailing-spaces)
 20:50     error    trailing spaces (trailing-spaces)
 21:46     error    trailing spaces (trailing-spaces)
 22:48     error    trailing spaces (trailing-spaces)
 36:51     error    trailing spaces (trailing-spaces)
 37:57     error    trailing spaces (trailing-spaces)
 38:52     error    trailing spaces (trailing-spaces)
 39:52     error    trailing spaces (trailing-spaces)
 40:55     error    trailing spaces (trailing-spaces)
 41:52     error    trailing spaces (trailing-spaces)
 42:58     error    trailing spaces (trailing-spaces)
 43:57     error    trailing spaces (trailing-spaces)
 44:57     error    trailing spaces (trailing-spaces)
 45:56     error    trailing spaces (trailing-spaces)
 46:54     error    trailing spaces (trailing-spaces)
```

```
47:59    error    trailing spaces  (trailing-spaces)
48:17    error    trailing spaces  (trailing-spaces)
61:51    error    trailing spaces  (trailing-spaces)
62:53    error    trailing spaces  (trailing-spaces)
63:52    error    trailing spaces  (trailing-spaces)
64:51    error    trailing spaces  (trailing-spaces)
65:54    error    trailing spaces  (trailing-spaces)
66:55    error    trailing spaces  (trailing-spaces)
67:60    error    trailing spaces  (trailing-spaces)
68:56    error    trailing spaces  (trailing-spaces)
69:52    error    trailing spaces  (trailing-spaces)
70:47    error    trailing spaces  (trailing-spaces)
76:43    error    trailing spaces  (trailing-spaces)
79:1     error    duplication of key "project" in mapping  (key-duplicates)
85:43    error    trailing spaces  (trailing-spaces)
87:51    error    trailing spaces  (trailing-spaces)
88:53    error    trailing spaces  (trailing-spaces)
89:52    error    trailing spaces  (trailing-spaces)
90:51    error    trailing spaces  (trailing-spaces)
91:54    error    trailing spaces  (trailing-spaces)
92:55    error    trailing spaces  (trailing-spaces)
93:60    error    trailing spaces  (trailing-spaces)
94:56    error    trailing spaces  (trailing-spaces)
95:52    error    trailing spaces  (trailing-spaces)
```

GPL Ghostscript 9.19: Unrecoverable error, exit code 1

=====  
Compliance Report  
=====

```
name: Ricky Carmickle
hid: 304
paper1: Nov 1 17 100%
paper2: 100%
project: Dec 04 17 100%
```

yamlcheck

---

wordcount

---

(null)

```
wc 304 project (null) 3600 report.tex  
wc 304 project (null) 67 report.pdf  
wc 304 project (null) 382 report.bib
```

```
find "
```

---

96: The initial question of "How far have spacewalks walked" is answered with the data generated here, with the caveat that input data is imperfect. The longest single spacewalk, in terms of time spent in EVA and distance traveled relative to the earth, is a 2001 Space Shuttle Mission working on the ISS where Jim Voss and Susan Helms worked outside for 8 hours and 56 minutes and traveled 246,946,540 meters relative to the earth's surface. Table \ref{t:mytable} includes the descriptive EVA data as well as the final calculated fields for the twenty most traveled spacewalks.

110: The geovisualization aspect of this project ultimately failed. After beginning the process of creating orbital ground traces through BaseMap, it was discovered that a precise ground trace requires some additional data either unavailable or unparsed from the Wikipedia data. A ground tracing function would require the time, to the second or minute, of both the beginning of the spacewalk and of the mission. The parameter fields recognized the original data parsing code reuse the field name "Date" for the multiple dates in the JSON tables on each mission page, including the date the page was published, which is irrelevant to this project. Further understanding of the JSON and html is needed before we could complete this step. Two other pieces of data needed for accuracy are the earth coordinates of the point over which the vessel first established an orbit with apoapsis and the decay rate for orbits which still experienced atmospheric drag.

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
6: \input{format/i523}
```

---

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

---

```
floats
```

---

- 96: The initial question of "How far have spacewalks walked" is answered with the data generated here, with the caveat that input data is imperfect. The longest single spacewalk, in terms of time spent in EVA and distance traveled relative to the earth, is a 2001 Space Shuttle Mission working on the ISS where Jim Voss and Susan Helms worked outside for 8 hours and 56 minutes and traveled 246,946,540 meters relative to the earth's surface. Table \ref{f:mytable} includes the descriptive EVA data as well as the final calculated fields for the twenty most traveled spacewalks.
- 98: Figure \ref{f:fig1} features the final calculated spacewalk distances for every spacewalk. Only the mission label is included. Repeats of a label indicate multiple EVAs by a mission's crew. The distances traveled range from the roughly 5,500 kilometers traversed by Alexei Leonov over 12 minutes in the very first human spacewalk to the quarter of a million kilometers traveled by Mission Specialists Jim Voss and Susan Helms in their 8 hour and 56 minute spacewalk as part of an early resupply mission to the ISS.
- 100: Figure \ref{f:fig2} is a parallel coordinate plot of the year a given EVA took place mapped to EVA duration. From this visualization, we see the increasing volume of spacewalks as well as the general increase in EVA time over the years as the technology to sustain low-earth orbit EVAs improved and space program efforts were focused on the construction of the ISS.
- 104: A glance at the orbital inclination reveals a distinct pattern in this category. Data points are clumped around the values of 28 and 50. Figure \ref{f:fig3} is a scatterplot of spacewalk duration and orbit inclination, colored by country \cite{Waskom2017}. There is a reason for the distinct clumping in the data along those values.
- 106: The inclination of an orbit is a measurement of the object's orbital plane compared to earth's equatorial plane. An orbiting object's inclination can never start at less than the latitude of the launch site. Lowering the inclination of an orbit below the starting latitude requires an adjustment burn at either orbital insertion or during orbit when latitude matches the desired inclination \cite{Davis2012}. In Figure \ref{f:fig3}, the lower

line of datapoints, at a value around 28 corresponds to the Latitude of Kennedy Space Center, which is the source of most American launches and is located at 2828N 8032W. The upper clumping of data points, at a value around 52 corresponds to the ISS inclination of 51.64 degrees.

- 108: The predicted and actual orbital period categories present an opportunity to test the validity of the calculations made with this dataset. We would expect an ideal 1-to-1 relationship between predicted and actual orbital periods. The predicted and actual periods are, in this instance, independent variables since predicted period was calculated with Apoapsis, Periapsis, semimajor axis, and constant values of earth's properties \cite{Wolfson2007,Muirden1982}. The actual orbit period measurement was not part of the calculation. The difference between these values has a mean of 6.265 seconds, a median of 0.210 seconds, and a standard deviation of 38.75. With single columns of mathematically independent variables, a linear regression is appropriate \cite{Community2017}. Figure \ref{f:fig4} displays the results of the linear regression model. The p-value of this model is 1.824 u-150. This is a demonstration of an effective relationship between variables as expected. In this model, the slop and intercept would carry some significance. The ideal relationship between two identical variables should have an intercept of 0 and a slope of 1 \cite{Frost2014}. The intercept of 441.38 is 7.996 percent of the mean of actual orbital periods. The slope of 0.9211 is 7.89 percent less than the expected slope. Both slope and intercept \cite{Frost2014} are almost precisely 8 percent deviated from the ideal expectation R-value and standard error confirm that the calculated orbital period is deviated by approximately 8 percent from the true values reported in the data \cite{Frost2014}.
- 119: In Figure \ref{f:fig1} we show a complete list of distance traveled in EVA by astronauts of every American or Russian mission launched until 2013.
- 122: \centering\includegraphics[height=1.0\textheight]{images/fig1.png}
- 123: \caption{All Distances}\label{f:fig1}
- 127: In Figure \ref{f:fig2} we show a parallel coordinate plot of mission launch year and EVA duration.
- 130: \centering\includegraphics[height=1.0\textheight]{images/fig2.png}
- 131: \caption{Parallel Coordinate Launch Year and EVA}\label{f:fig2}
- 135: In Figure \ref{f:fig3} we show a scatterplot of mission inclination colored by country of launch. A pattern reflective of launchsites and mission destinations is apparent.
- 137: \begin{figure}[p]

```
138: \centering\includegraphics[width=\columnwidth]{images/fig3.png}
139: \caption{Inclination and Country}\label{f:fig3}
143: In Figure \ref{f:fig4} we show the linear regression results of
    the reported orbital period versus the predicted period
    calculated from other fields in the data.
145: \begin{figure}[htb]
146: \centering\includegraphics[width=\columnwidth]{images/fig4.png}
147: \caption{Linear Regression}\label{f:fig4}
157: \label{mytable}
```

figures 2  
tables 0  
includegraphics 4  
labels 5  
refs 10  
floats 2

False : ref check passed: (refs >= figures + tables)  
False : label check passed: (refs >= figures + tables)  
False : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)

Label/ref check  
155: \caption{Table 1. Longest Distance Spacewalks}  
passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

```
158: \resizebox{\textwidth}{!}{%
```

passed: False

below\_check

---

bibtex

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--can't use both author and editor fields in Muirden1982
Warning--empty address in Muirden1982
Warning--can't use both author and editor fields in Wolfson2007
Warning--empty address in Wolfson2007
(There were 4 warnings)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

```
non ascii found 8217
non ascii found 8217
non ascii found 8216
non ascii found 8217
non ascii found 8216
non ascii found 8217
non ascii found 8216
non ascii found 8217
```

```
non ascii found 8216
non ascii found 8217
non ascii found 8220
non ascii found 8221
non ascii found 8220
non ascii found 8221
non ascii found 8217
non ascii found 8220
non ascii found 8221
non ascii found 8220
non ascii found 8221
non ascii found 8216
non ascii found 8217
non ascii found 8216
non ascii found 8217
non ascii found 8216
non ascii found 8217
non ascii found 8217
non ascii found 176
non ascii found 8242
non ascii found 176
non ascii found 8242
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

```
latex report
=====
```

```
[2017-12-16 09.35.46] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.2    L95    : 't:mytable' undefined
```





```
16:48    error    trailing spaces  (trailing-spaces)
17:49    error    trailing spaces  (trailing-spaces)
18:51    error    trailing spaces  (trailing-spaces)
19:45    error    trailing spaces  (trailing-spaces)
20:50    error    trailing spaces  (trailing-spaces)
21:46    error    trailing spaces  (trailing-spaces)
22:48    error    trailing spaces  (trailing-spaces)
36:51    error    trailing spaces  (trailing-spaces)
37:57    error    trailing spaces  (trailing-spaces)
38:52    error    trailing spaces  (trailing-spaces)
39:52    error    trailing spaces  (trailing-spaces)
40:55    error    trailing spaces  (trailing-spaces)
41:52    error    trailing spaces  (trailing-spaces)
42:58    error    trailing spaces  (trailing-spaces)
43:57    error    trailing spaces  (trailing-spaces)
44:57    error    trailing spaces  (trailing-spaces)
45:56    error    trailing spaces  (trailing-spaces)
46:54    error    trailing spaces  (trailing-spaces)
47:59    error    trailing spaces  (trailing-spaces)
48:17    error    trailing spaces  (trailing-spaces)
61:51    error    trailing spaces  (trailing-spaces)
62:53    error    trailing spaces  (trailing-spaces)
63:52    error    trailing spaces  (trailing-spaces)
64:51    error    trailing spaces  (trailing-spaces)
65:54    error    trailing spaces  (trailing-spaces)
66:55    error    trailing spaces  (trailing-spaces)
67:60    error    trailing spaces  (trailing-spaces)
68:56    error    trailing spaces  (trailing-spaces)
69:52    error    trailing spaces  (trailing-spaces)
70:47    error    trailing spaces  (trailing-spaces)
76:43    error    trailing spaces  (trailing-spaces)
79:1     error    duplication of key "project" in mapping  (key-duplicates)
85:43    error    trailing spaces  (trailing-spaces)
87:51    error    trailing spaces  (trailing-spaces)
88:53    error    trailing spaces  (trailing-spaces)
89:52    error    trailing spaces  (trailing-spaces)
90:51    error    trailing spaces  (trailing-spaces)
91:54    error    trailing spaces  (trailing-spaces)
92:55    error    trailing spaces  (trailing-spaces)
93:60    error    trailing spaces  (trailing-spaces)
94:56    error    trailing spaces  (trailing-spaces)
95:52    error    trailing spaces  (trailing-spaces)
```

```
=====
name: Ricky Carmickle
hid: 304
paper1: Nov 1 17 100%
paper2: 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 304 project (null) 3600 report.tex
wc 304 project (null) 3539 report.pdf
wc 304 project (null) 382 report.bib
```

```
find "
```

---

96: The initial question of "How far have spacewalks walked" is answered with the data generated here, with the caveat that input data is imperfect. The longest single spacewalk, in terms of time spent in EVA and distance traveled relative to the earth, is a 2001 Space Shuttle Mission working on the ISS where Jim Voss and Susan Helms worked outside for 8 hours and 56 minutes and traveled 246,946,540 meters relative to the earth's surface. Table \ref{mytable} includes the descriptive EVA data as well as the final calculated fields for the twenty most traveled spacewalks.

110: The geovisualization aspect of this project ultimately failed. After beginning the process of creating orbital ground traces through BaseMap, it was discovered that a precise ground trace requires some additional data either unavailable or unparsed from the Wikipedia data. A ground tracing function would require the time, to the second or minute, of both the beginning of the spacewalk and of the mission. The parameter fields recognized the original data parsing code reuse the field name "Date" for the multiple dates in the JSON tables on each mission page, including the date the page was published, which is irrelevant to this project. Further understanding of the JSON and html is needed before we could complete this step. Two other pieces of data

needed for accuracy are the earth coordinates of the point over which the vessel first established an orbit with apoapsis and the decay rate for orbits which still experienced atmospheric drag.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

6: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

96: The initial question of "How far have spacewalks walked" is answered with the data generated here, with the caveat that input data is imperfect. The longest single spacewalk, in terms of time spent in EVA and distance traveled relative to the earth, is a 2001 Space Shuttle Mission working on the ISS where Jim Voss and Susan Helms worked outside for 8 hours and 56 minutes and traveled 246,946,540 meters relative to the earth's surface. Table \ref{t:mytable} includes the descriptive EVA data as well as the final calculated fields for the twenty most traveled spacewalks.

98: Figure \ref{f:fig1} features the final calculated spacewalk distances for every spacewalk. Only the mission label is included. Repeats of a label indicate multiple EVAs by a mission's crew. The distances traveled range from the roughly 5,500 kilometers traversed by Alexei Leonov over 12 minutes in the very first human spacewalk to the quarter of a million kilometers traveled by Mission Specialists Jim Voss and Susan Helms in their 8 hour and 56 minute spacewalk as part of an early resupply mission to the ISS.

100: Figure \ref{f:fig2} is a parallel coordinate plot of the year a given EVA took place mapped to EVA duration. From this

visualization, we see the increasing volume of spacewalks as well as the general increase in EVA time over the years as the technology to sustain low-earth orbit EVAs improved and space program efforts were focused on the construction of the ISS.

- 104: A glance at the orbital inclination reveals a distinct pattern in this category. Data points are clumped around the values of 28 and 50. Figure \ref{f:fig3} is a scatterplot of spacewalk duration and orbit inclination, colored by country \cite{Waskom2017}. There is a reason for the distinct clumping in the data along those values.
- 106: The inclination of an orbit is a measurement of the object's orbital plane compared to earth's equatorial plane. An orbiting object's inclination can never start at less than the latitude of the launch site. Lowering the inclination of an orbit below the starting latitude requires an adjustment burn at either orbital insertion or during orbit when latitude matches the desired inclination \cite{Davis2012}. In Figure \ref{f:fig3}, the lower line of datapoints, at a value around 28 corresponds to the Latitude of Kennedy Space Center, which is the source of most American launches and is located at 2828N 8032W. The upper clumping of data points, at a value around 52 corresponds to the ISS inclination of 51.64 degrees.
- 108: The predicted and actual orbital period categories present an opportunity to test the validity of the calculations made with this dataset. We would expect an ideal 1-to-1 relationship between predicted and actual orbital periods. The predicted and actual periods are, in this instance, independent variables since predicted period was calculated with Apoapsis, Periapsis, semimajor axis, and constant values of earth's properties \cite{Wolfson2007,Muirden1982}. The actual orbit period measurement was not part of the calculation. The difference between these values has a mean of 6.265 seconds, a median of 0.210 seconds, and a standard deviation of 38.75. With single columns of mathematically independent variables, a linear regression is appropriate \cite{Community2017}. Figure \ref{f:fig4} displays the results of the linear regression model. The p-value of this model is 1.824 u-150. This is a demonstration of an effective relationship between variables as expected. In this model, the slope and intercept would carry some significance. The ideal relationship between two identical variables should have an intercept of 0 and a slope of 1 \cite{Frost2014}. The intercept of 441.38 is 7.996 percent of the mean of actual orbital periods. The slope of 0.9211 is 7.89 percent less than the expected slope. Both slope and intercept \cite{Frost2014} are almost precisely 8 percent deviated from the ideal expectation R-value and standard error confirm that the calculated orbital

period is deviated by approximately 8 percent from the true values reported in the data \cite{Frost2014}.

119: In Figure \ref{f:fig1} we show a complete list of distance traveled in EVA by astronauts of every American or Russian mission launched until 2013.

122: \centering\includegraphics[height=1.0\textheight]{images/fig1.png}

123: \caption{All Distances}\label{f:fig1}

127: In Figure \ref{f:fig2} we show a parallel coordinate plot of mission launch year and EVA duration.

130: \centering\includegraphics[height=1.0\textheight]{images/fig2.png}

131: \caption{Parallel Coordinate Launch Year and EVA}\label{f:fig2}

135: In Figure \ref{f:fig3} we show a scatterplot of mission inclination colored by country of launch. A pattern reflective of launchsites and mission destinations is apparent.

137: \begin{figure}[p]

138: \centering\includegraphics[width=\columnwidth]{images/fig3.png}

139: \caption{Inclination and Country}\label{f:fig3}

143: In Figure \ref{f:fig4} we show the linear regression results of the reported orbital period versus the predicted period calculated from other fields in the data.

145: \begin{figure}[htb]

146: \centering\includegraphics[width=\columnwidth]{images/fig4.png}

147: \caption{Linear Regression}\label{f:fig4}

157: \label{mytable}

figures 2  
 tables 0  
 includegraphics 4  
 labels 5  
 refs 10  
 floats 2

False : ref check passed: (refs >= figures + tables)  
 False : label check passed: (refs >= figures + tables)  
 False : include graphics passed: (figures >= includegraphics)  
 True : check if all figures are referred to: (refs >= labels)

Label/ref check  
 155: \caption{Table 1. Longest Distance Spacewalks}  
 passed: False -> labels or refs used wrong

When using figures use columnwidth  
 [width=1.0\columnwidth]  
 do not change the number to a smaller fraction

```
find textwidth
```

---

```
158: \resizebox{\textwidth}{!}{%
```

```
passed: False
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
```

```
The top-level auxiliary file: report.aux
```

```
The style file: ACM-Reference-Format.bst
```

```
Database file #1: report.bib
```

```
Warning--can't use both author and editor fields in Muirden1982
```

```
Warning--empty address in Muirden1982
```

```
Warning--can't use both author and editor fields in Wolfson2007
```

```
Warning--empty address in Wolfson2007
```

```
(There were 4 warnings)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

non ascii found 8217  
non ascii found 8217  
non ascii found 8216  
non ascii found 8217  
non ascii found 8216  
non ascii found 8217  
non ascii found 8216  
non ascii found 8217  
non ascii found 8216  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8220  
non ascii found 8221  
non ascii found 8216  
non ascii found 8217  
non ascii found 8216  
non ascii found 8217  
non ascii found 176  
non ascii found 8242  
non ascii found 176  
non ascii found 8242

The following tests are optional

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

# Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks

Neha Rawat  
Indiana University  
Bloomington, Indiana  
nrawat@iu.edu

## ABSTRACT

With the increase in internet traffic, threats on the network have also increased. Denial-of-service attacks are cyber attacks wherein a perpetrator, due to any kind of malicious intent, tries to make a resource on the network unavailable to its intended users and carries it out by swamping the system or resource with excess requests in order to overload it and prevent users from accessing it. A much more dangerous variety of such an attack is if it is distributed i.e. coming from various sources. Big Data analytics, however, can be used to detect such attacks by having the ability to store the voluminous logs of such attacks and using the data and machine learning techniques to design an anomaly detection system (using a classification model) to detect and prevent these attacks. This project will aim to explore such classification models, design and train the most optimum model and display its effects using a DDoS network traffic logs dataset.

## KEYWORDS

i523, HID224, Denial-of-Service, Intrusion Detection, KDD Cup'99 dataset, Machine Learning, Apache Spark

## 1 INTRODUCTION

The Internet allows us several comforts and functionalities in our day-to-day lives. With the increasing flexibility and accessibility provided by technology, the Internet has become an indispensable part of our life. However, this same accessibility often provides openings for malicious attackers to enter. Security over the Internet is an interdependent factor, with the security of one user depending on rest of the global network [1]. Denial-of-Service attacks are attacks by such malicious users in order to disrupt the accessibility of other legitimate users to a Web Service or application [7]. The objectives of such attacks are mainly malicious, driven out of revenge or for some material gain. The attacks seriously hinder the productivity of the victim, as the resources available are not sufficient to handle the oncoming flood of requests. This attack increases in complexity when there are multiple sources of attacks, resulting in a Distributed Denial-of-Service attack. “In the case of a Distributed Denial-of-Service (DDoS) attack, an attacker uses multiple sources - which may be compromised or controlled by a group of collaborators - to orchestrate an attack against a target” [7]. A small batch of requests sent by an attacker may be enough to generate a large amount of unwanted traffic. The earliest of these attacks was when a DDoS tool called Trinoo, deployed in at least 227 systems, flooded a University of Minnesota computer, which was subsequently rendered useless for more than two days [1]. Figure 1 shows how a Distributed Denial-of-Service attack occurs.

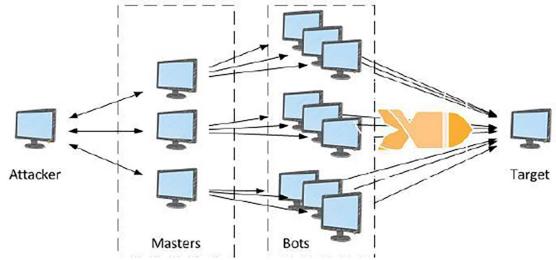


Figure 1: Distributed Denial-of-Service Attack [7]

As the connectivity increases in our everyday lives, so have the risks for DDoS attacks. The Internet of Things (IoT) for example, has opened up a whole new avenue for Denial-of-Service attackers. Earlier, limited to attacks over the Internet which mostly affected a user’s computer, with the advent of IoT, the scope of attacks on other smart devices has increased considerably. These devices could be used as pawns in a DDoS attack network and could even be the intended targets for such an attack. Some of the largest DDoS attacks till date are as given: In March of 2013, the DDoS attack on Spamhaus saw 120 Gbps of traffic on their network, in August of 2013, a “part of the Chinese internet went down in one of the largest DDoS attacks”, in the Spring of 2015, UK-based phone carrier Carphone Warehouse got attacked and hackers stole millions of customers’ data and in January of 2016, some HSBC customers were inhibited from accessing their online banking accounts, which caused a great upheaval as it was “two days before the tax payment deadline in the United Kingdom” [11]. We can see that these attacks, if allowed to happen, have great damage potential. Hence, DDoS mitigation service providers like Imperva Incapsula Enterprise, Arbor Cloud, Verisign, DOSarrest and CloudFlare, have their work cut out for them to detect and prevent such attacks, which are increasing in their reach and complexity [10].

## 2 DDoS ATTACK TYPES AND ARCHITECTURE

In order to prevent a DDoS attack, it is important to know the points in a network where the attack is expected to occur and the type of attack that can occur. Referring to an Open Systems Interconnection(OSI) model, we can usually narrow down the layers which could be affected by a potential attack to the Network, Transport, Presentation and Application layers [7]. Figure 2 shows an Open Systems Interconnection Model with the layers highlighted where DDoS attacks are most common.

#	Layer	Unit	Description	Vector Examples
7	Application	Data	Network process to application	HTTP floods, DNS query floods
6	Presentation	Data	Data representation and encryption	SSL abuse
5	Session	Data	Interhost communication	N/A
4	Transport	Segments	End-to-end connections and reliability	SYN floods
3	Network	Packets	Path determination and logical addressing	UDP reflection attacks
2	Data Link	Frames	Physical addressing	N/A
1	Physical	Bits	Media, signal, and binary transmission	N/A

Figure 2: Open Systems Interconnection Model [7]

Apart from this, the DDoS attacks generally have a specific architecture and follow certain strategies. Knowledge of the pathway which a Denial-of-Service attack follows is essential to detecting and mitigating it.

## 2.1 DDoS Attack Types

The DDoS attacks in the Network and Transport layers are generally of the User Datagram Protocol (UDP) reflection and synchronize (SYN) flood types [7]. The UDP protocol can allow the attacker to fake the source of a request sent to a server and generate a larger response. The amplification factor of a protocol (request to response size) will result in an overwhelming response to a comparatively smaller request. “For example, the amplification factor for DNS can be in the 28 to 54 range - which means an attacker can send a request payload of 64 bytes to a DNS server and generate over 3400 bytes of unwanted traffic” [7]. A SYN flood attack is based on employing all the resources of a system and exhausting them by leaving connections half-open. For example, when an user connects to a TCP service, the client will send a SYN packet and the server will return a SYN-ACK, expecting the client to return an ACK and completing the handshake. In a SYN flood attack, the ACK is not returned and so the server is stuck in this state which prevents other users from connecting to it [7].

In the Presentation and Application layers, the DDoS attacks are slightly different. The most common of such attacks are “HTTP floods, cache-busting attacks, and WordPress XML-RPC floods” [7]. In an HTTP flood attack, the attacker sends HTTP requests under the guise of a real user or web service. These attacks target a resource or try to emulate human behavior. Cache-busting attacks are a specialized version of HTTP flood attacks that use “variations in the query string to circumvent content delivery network (CDN) caching which results in origin fetches, causing additional strain on the origin web server” [7]. A WordPress XML-RPC flood (WordPress pingback flood) is used by an attacker to misuse the XML-RPC API function of a website hosted on WordPress software to generate HTTP flood requests. This type of attack has *WordPress* present in the HTTP request header and so is clearly recognizable [7].

## 2.2 DDoS Attack Architecture

“DDoS attack networks follow two types of architectures: the Agent-Handler architecture and the Internet Relay Chat (IRC)-based architecture” [1]. The components of an Agent-Handler architecture are clients, handlers, and agents. In this type of architecture, the attacker connects with the rest of the attack system at the client point. The handlers are generally software packages available over

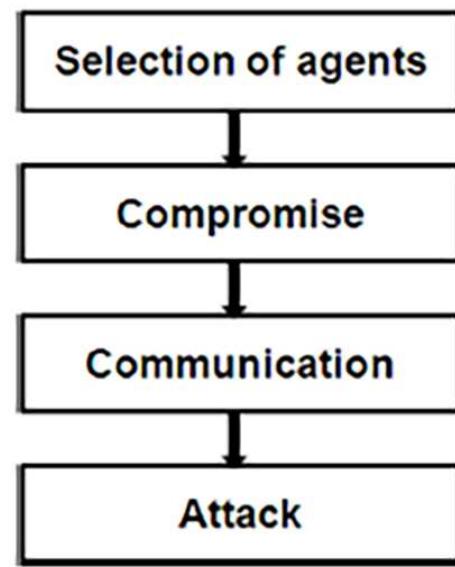
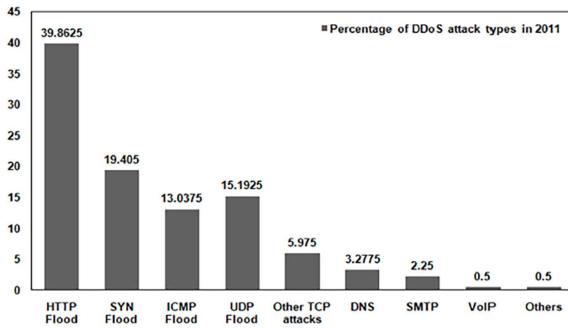


Figure 3: Steps of a Denial-of-Service attack [1]

the Internet which are used by the client to connect to the agents. The agent softwares are placed in the vulnerable systems that are finally used to implement the attack. Often, the users of the agent systems are not aware of the attack being carried out [1]. In the IRC-based architecture, “an IRC communication channel is used to connect the client(s) to the agents” [1]. IRC ports are employed to send commands to the agents, making the DDoS command packets harder to trace (as these channels have a lot of traffic) [1]. When launching a DDoS attack, the attacker goes through some steps common to both types of architectures [1]. First, the attacker tries to identify vulnerable systems that can be used as agents. The resources of these systems are used to generate a powerful attack stream. Next, the attacker plants the handler software code in the compromised system and ensures steps to prevent the code from being detected. These compromised systems are often referred to as *zombies*. Sometimes, the attacker creates several intermediate layers between the *zombies* and the victim to hinder traceability. Thirdly, the attacker communicates with the handler codes placed via protocols like TCP or UDP, and decides the scheduling of the attacks. Post the complete setup, the attacker launches the attack on the victim’s machine or server and renders it unusable [1]. In an IRC-based architecture, most of the above steps remain same, but an IRC-channel is used for communication purposes. This helps the attacker as even if one *zombie* or *bot* is discovered, the identities of the others is still hidden, as IRC-channels are difficult to detect [1]. Figure 3 shows the steps of a Denial-of-Service attack execution.



**Figure 4: Different Denial-of-Service attack type statistics [1]**

### 3 DDOS ATTACK DEFENSE METHODOLOGIES

In the previous section, we explored the common types of DDoS attacks and the general architecture that they follow. These different types of attacks are used with variation by attackers in their attempts to obstruct utilization of resources. Figure 4 shows the percentage of different Denial-of-Service attacks in 2011 by type. The different types of DDoS attacks and their improvement throughout time has also invoked different defense mechanisms against these attacks. DDoS defense mechanisms are usually employed at three points in the attack network : Victim-end, Source-end and Intermediate-Network [1]. Victim-end detection approaches are generally incorporated in the routers of victim networks. A detection system is used to detect intrusion based on different techniques. Detecting DDoS attacks at this point is relatively easy and the most practically applicable, but has the disadvantage of detection only after the attack has reached the victim and legitimate users have already been denied services [1]. Source-end detection system works similarly to the victim-end detection system apart from “a throttling component”, which is added to force a rate limit on outgoing connections. The detection system then compares both incoming and outgoing network traffic with normal traffic benchmarks to detect an attack. This is probably the ideal defense mechanism, but faces challenges in the deployment of a detection system at the source and difficulty in identification in case of multiple sources [1]. The intermediate-network defense mechanism acts like a middle-ground between the victim-end and source-end systems. It acts like a collaborative model which depends upon communication and sharing of information between all routers on the network. Hence, this too suffers from the problem of deployability, as even one router missing on the network could hinder the traceback process [1].

From the above defense mechanism schemes, we can garner that detection of these attacks forms a major part of the preventive process. The most commonly used detection methodologies for defense against DDoS are as follows: Statistical Methods, Soft-Computing and Machine Learning Methods and Knowledge-Based Methods [1].

### 3.1 Statistical Methods

Statistical Methods follow the statistical properties of the distribution of incoming and outgoing network traffic for detection of DDoS attacks. The distributions (or statistical estimates generated using it) are compared with those for a normal traffic signature. An example of the same is the use of cumulative deviation from normal to detect DDoS attacks. Similarly, a periodic deviation analysis from the normal pattern can be used to detect intrusions [1]. Another example, is the use of a two-sample t-test to detect DDoS signatures by comparing the SYN arrival rate distribution with the distribution of a normal SYN arrival rate (after confirming a gaussian distribution for it). If the difference is considered significant according to the t-test, the traffic is marked as potentially containing attack packets [1]. A prediction method designed by Zhang et al. [12] uses an Auto Regressive Integrated Auto Regressive (ARIMA) model for their detection system.

### 3.2 Soft-Computing and Machine Learning Methods

The voluminous network traffic data generated can be leveraged by a soft-computing system like a neural network or a data mining/machine learning model to design a classifier that differentiates between normal traffic and intrusions. An example is the use of statistical preprocessing for extraction of relevant features from the traffic followed by an unsupervised neural net to classify traffic signatures as either a DDoS attack or normal [1]. Another case is the use of a Radial Basis Function (RBF) neural network to analyze attack packets and classify them as normal or harmful [1]. Machine learning algorithms like K-Nearest Neighbors and Support Vector Machines can be used as excellent classifiers for incoming network traffic. Fuzzy networks can also be used in the decision-making process while separating normal traffic packets from potentially harmful ones [1].

### 3.3 Knowledge-Based Methods

In knowledge-based methods, network traffic features are compared with predefined patterns of attack. Some examples of knowledge-based methodologies include “expert systems, signature analysis, self organizing maps, and state transition analysis” [1]. Heuristics can be used to analyze traffic characteristics and classify them as DDoS or otherwise. An excellent example is that of a DDoS detection system which used a “gossip based communication mechanism” to exchange information about network attacks among independent detection nodes in order to use the aggregate data to identify network attacks [1]. Another model, used temporal-correlation based method to extract features and spatial-correlation for detection to correctly identify DDoS attacks [1].

## 4 DDOS ATTACK DETECTION MODEL

For this project, we have worked on the design and implementation of an optimal DDoS detection model (based on Soft-Computing and Machine Learning algorithms) by training and implementing several potential models and creating an ensemble model from the best ones. We have also explored the traffic logs dataset to identify patterns via unsupervised means.

## 4.1 Data Description

The KDD Cup'99 dataset [6] has been used for our data analysis. This dataset has been derived from the 1998 DARPA Intrusion Detection Evaluation Program dataset [8] which was prepared and managed by MIT Lincoln Labs. The data was simulated to evaluate study in intrusion detection. It comprises of a “wide variety of intrusions simulated in a military network environment” [6]. The original data comprised of around five million records. Hence, we use a 10 percent subset of the original train and test datasets for our analysis purposes.

## 4.2 Data Exploration and Processing

The data exploration and analysis for this project has been implemented using Python on *Jupyter Notebook*. The *Jupyter Notebook* provides us with “an open-source web application that allows us to create and share documents that contain live code, equations, visualizations and narrative text” [5].

For data loading, we use the *Pandas* library in python, which is one of the largest and most flexible data managing libraries and offers a wide variety of options for data handling and manipulation using data frames. After loading the datasets, we explore some of the features of the dataset. From the documentation on the KDD Cup'99 dataset, we know that the data consists of a wide variety of network attacks, but the five main classes of network traffic are as follows: normal (normal network traffic), DoS/DDoS (Denial-of-Service network traffic), R2L (unauthorized access from a remote machine traffic), U2R (unauthorized access to local superuser privileges traffic) and probing [6]. Also, the test dataset consists of an additional 14 attack types which are not present in the training data. However, these new attack types are also a part of the above five categories and the purpose behind their addition in the dataset was to prove that new variants can also be detected using signatures of the preexisting types of attacks [6].

For plotting and visualization purposes, we use *Matplotlib* and *Seaborn* - two excellent visualization libraries offered by Python. First, we check for nulls in the train and test dataset, but find none. Secondly, we check the three categorical columns in the data, to ensure same levels in both the training and test dataset. We find that the training dataset has an additional level in the *service* column. For simplicity, we remove the categorical columns from our analysis dataset and continue our work on only the numerical columns. We now explore the target label column which specifies the *attack type* or the network traffic class. We map the labels to five core categories discussed previously and compare them for the training and testing set. Figure 5 shows the Attack Type distribution in the training and test datasets

We can observe that DoS attacks form the majority of all the attack types (98.67 percent out of all attacks in training set; 91.78 percent out of all attacks in test set). Hence, we broadly classify the target labels as *normal* and *bad* for intrusion detection. We also include the individual labels for the multi-label classification part.

Post this, we create pair plots for the first few variables in order to view individual distributions as well as correlations. Figure 6 shows the pair plot between the first 15 variables in the training dataset. We observe that the data seems to be skewed, indicating the need for standardizing the features. Also, there do not seem to

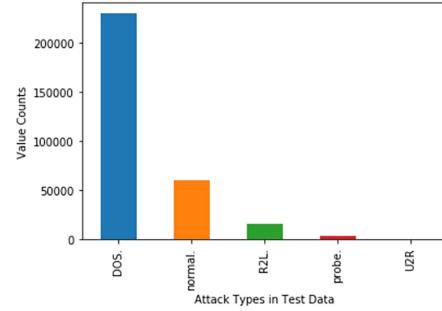
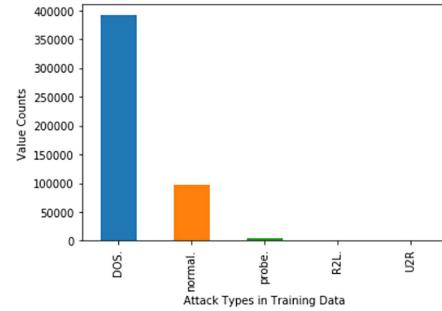


Figure 5: Attack Type Distributions

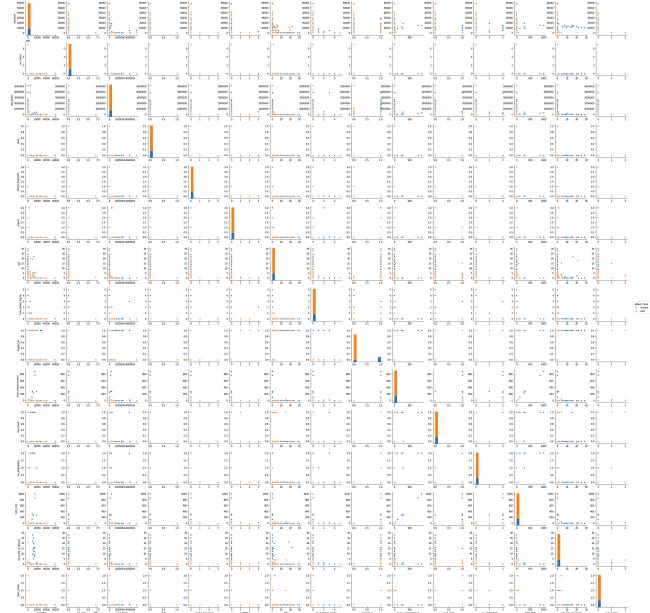


Figure 6: Pair plot for Training Features

be a lot of correlated variables in the dataset.

We proceed with separating the binary variables (mentioned in the documentation) from the continuous variables and scaling the continuous variables using mean normalization in the training dataset. We then apply the same transformations to the test dataset. Post

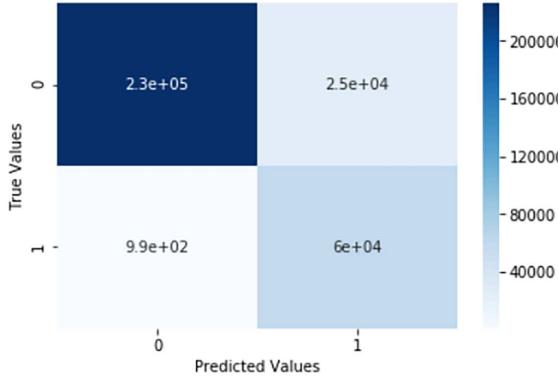


Figure 7: Logistic Regression Confusion Matrix - 2-class classification

this, we consolidate all our features and get the final processed datasets for training and testing.

### 4.3 Data Analysis

Once we are ready with our final datasets, we design the required detection models by training them on the training data and testing their performance on the test data. For the design of the models, we use the *scikit-learn* or *sklearn* package in python, which contains a plethora of resources for statistical and machine learning methodologies. For most models, we also employ the *n\_jobs* parameter present in the models for parallelization purposes [9]. For performance tests, we calculate the accuracy, precision, recall and F1 score for the model (for both 2-label and multi-label classification). The confusion matrix generated in each case displays the classes as follows: 2-class classification (0 - bad, 1 - normal) ; Multi-class classification (0 - Dos/DDoS, 1 - R2L, 2 - U2R, 3 - normal, 4 - probe).

**4.3.1 Logistic Regression.** Logistic Regression is a machine learning algorithm based on the regression model which is used to fit a model to describe the relationship between a dependent (categorical target) and one or more independent variables. Used mainly for classification purposes, the target variable in a logistic regression model is mainly binary, although the method can be used for multi-class classification too. The basis of logistic regression is a *logistic function* (usually a sigmoid function) which keeps the output values bounded between 0 and 1. This function is fit using a *maximum likelihood* methodology which attempts to estimate the coefficients of the regression equation such that the probability outputs match as closely as possible to the true output values [4].

We train two logistic regression models - one for the 2-class classification and one for the multi-class classification. The model for the 2-class classification is fit as per the default parameters, with the regularization parameter as 0.01 for stronger regularization. For the multi-class classification (since this is not the default type for a logistic regression model), we use a specific solver method known as *Stochastic Average Gradient Descent Solver* [9]. Figure 7 shows the 2-class confusion matrix for logistic regression. Figure 8 shows the multi-class confusion matrix for logistic regression.

The overall accuracy, recall, precision and F1 score for the 2-class

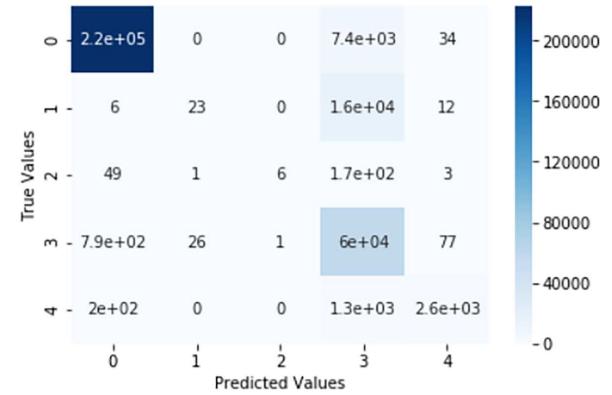


Figure 8: Logistic Regression Confusion Matrix - Multi-class classification

classification are as follows: 91.7, 94.2, 85.0 and 88.3 percent. The same for the multi-class classification are as follows: 91.5, 52.2, 79.4 and 52.3. We can observe that the accuracy of the model seems to be good for the 2-class classification but the recall and F1 scores decrease for the multi-class classification (due to the decrease in recall for the U2R and R2L classes, which have a higher proportion in test as compared to train data).

**4.3.2 K-Nearest Neighbors.** The K-Nearest Neighbors algorithm selects the  $k$  nearest points to the test data point (depending upon a predefined distance metric), present in the training data point, and assign it the class label depending on the majority class label present among the  $k$  training data points. Being a non-parametric method, KNN does not assume any initial distribution or form of data [4].

We train two KNN ( $k=5$ ) models - one for the 2-class classification and one for the multi-class classification. For both the classification models, instead of taking the *brute force* or traditional approach, we use an optimized approach known as *Ball Tree Method* [9], which is a tree based method that endeavors to reduce the number of distance computations by encoding the distance information more efficiently. It recursively divides the data according to a hyper-sphere determined by a particular centroid and radius, and reduces the participants for a neighbor search using triangle inequality [9]. This method works better for data in high dimensions (similar to the dataset for our analysis). Also, we take the distance metric as Manhattan Distance instead of the commonly used Euclidean Distance metric, due to better properties of Manhattan distance in higher dimensions.

Figure 9 shows the 2-class confusion matrix for KNN. Figure 10 shows the multi-class confusion matrix for KNN. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.35, 94.64, 85.95 and 89.20 percent. The same for the multi-class classification are as follows: 92.08, 55.90, 80.16 and 55.17. We can observe that the accuracy of the model increases as compared to a simple logistic regression model for the 2-class classification. The recall and F1 scores too increase for the multi-class classification case.

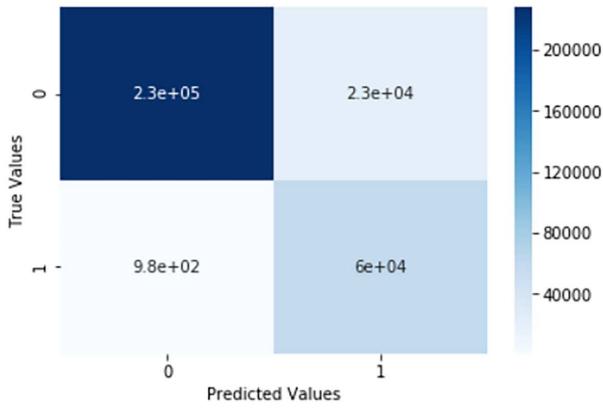


Figure 9: KNN Confusion Matrix - 2-class classification

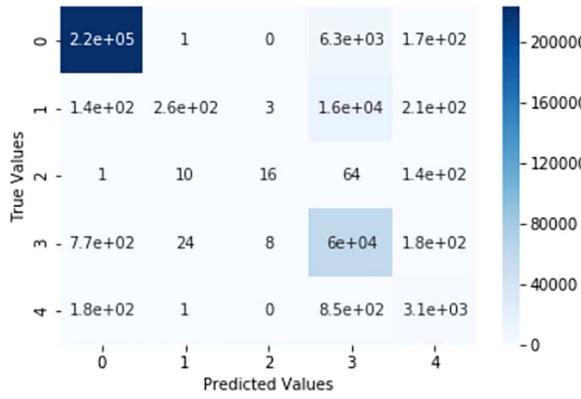


Figure 10: KNN Confusion Matrix - Multi-class classification

**4.3.3 Support Vector Machine - Linear.** A Support Vector Machine is a model based on the maximal margin classifier i.e. classification based on an optimal separating hyperplane. The support vector machine extends this concept further and to non-linear decision boundaries as well. It uses a function referred to as the *kernel* which acts as quantification of the similarity between observations. Therefore, for non-linear cases a variety of kernels such as radial or polynomial can be employed for classification purposes [4].

We train two linear SVM models - one for the 2-class classification and one for the multi-class classification. We implement this classifier using a *Bagging Classifier* which uses the base SVM classifier on different subsets of data drawn with replacement (also referred to as bootstrapping) and aggregates the results to given the final output [9]. Figure 11 shows the 2-class confusion matrix for linear SVM. Figure 12 shows the multi-class confusion matrix for linear SVM.

The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.23, 94.55, 85.78 and 89.04 percent. The same for the multi-class classification are as follows: 89.14, 54.91, 83.44 and 55.81. We can observe that the accuracy of the model increases as compared to a simple logistic regression model

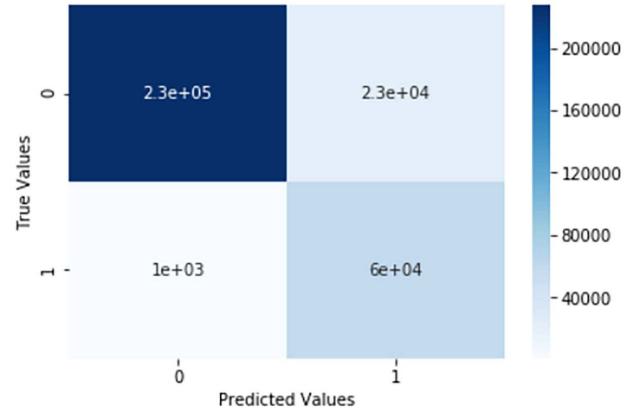


Figure 11: Linear SVM Confusion Matrix - 2-class classification

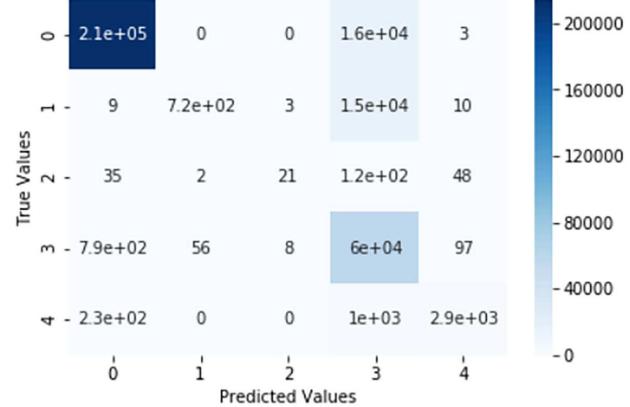
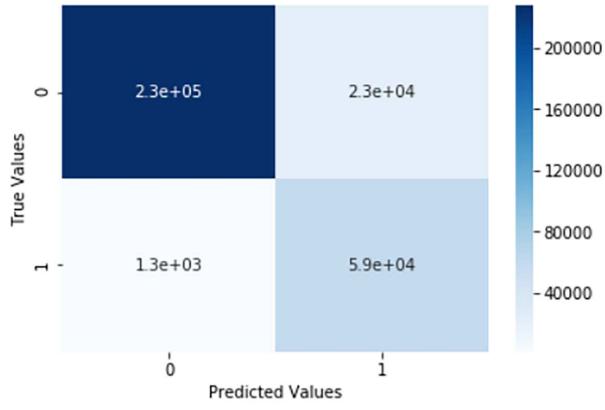


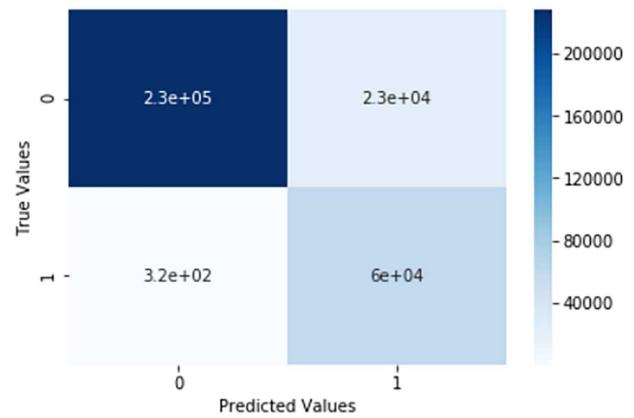
Figure 12: Linear SVM Confusion Matrix - Multi-class classification

but is lower than the KNN model for the 2-class classification. The recall and F1 scores too increase compared to logistic regression but are similar to that of KNN for the multi-class classification case.

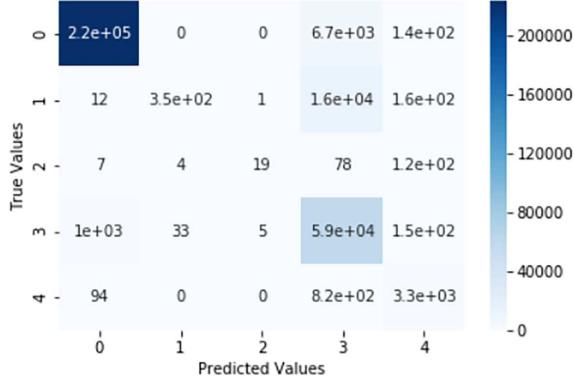
**4.3.4 Support Vector Machine - Polynomial.** Here, we train two SVM models (with polynomial kernels of degree=3) and using a *Bagging Classifier* - one for the 2-class classification and one for the multi-class classification. Figure 13 shows the 2-class confusion matrix for polynomial SVM. Figure 14 shows the multi-class confusion matrix for polynomial SVM. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.28, 94.41, 85.87 and 89.08 percent. The same for the multi-class classification are as follows: 91.95, 56.74, 84.60 and 56.38. We can observe that the accuracy of this model too is lower than the KNN model for the 2-class classification. However, the recall and F1 scores are higher than KNN too (correctly classifies more DoS/DDoS and probe attacks than linear



**Figure 13: Polynomial SVM Confusion Matrix - 2-class classification**



**Figure 15: Random Forest Confusion Matrix - 2-class classification**



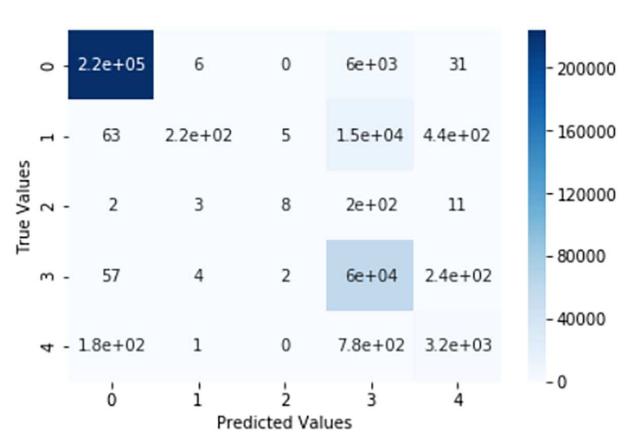
**Figure 14: Polynomial SVM Confusion Matrix - Multi-class classification**

SVM and more R2L and probe attacks than KNN) for the multi-class classification case. Overall, the performance is similar to KNN.

**4.3.5 Random Forest.** A random forest model works as an improvement over individual decision trees through building a number of decision trees on bootstrapped samples along with decorrelating the individual trees by choosing only a random subset of predictors out of the total predictors while constructing trees. At each split, a fresh subset of predictors is used which implements the decorrelation of features [4].

We train two random forest models - one for the 2-class classification and one for the multi-class classification. The selection of the subset of features is taken as the default parameter i.e. square root of the total number of features [9]. Figure 15 shows the 2-class confusion matrix for a Random Forest. Figure 16 shows the multi-class confusion matrix for a Random Forest.

The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.64, 95.22, 86.31, 89.63 percent. The same for the multi-class classification are as follows: 92.44, 55.74, 80.39 and 54.26. We can observe that the accuracy of this model



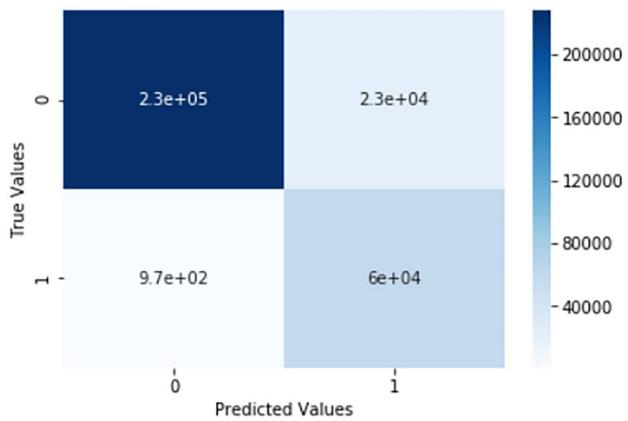
**Figure 16: Random Forest Confusion Matrix - Multi-class classification**

higher than all the previous models for the 2-class classification. The recall and F1 score for multi-class classification is comparable to the SVM models.

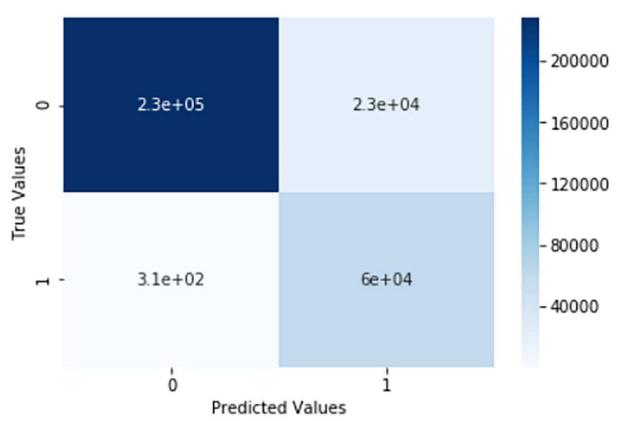
**4.3.6 Neural Networks : Multi-Layer Perceptron.** Neural Networks are soft-computing techniques that attempt to replicate information processing in biological systems, and thus have excellent learning capabilities. When used for pattern recognition or classification purposes, the most useful Neural Network is that of Multi-Layer Perceptron which basically acts as multiple layers of logistic regression models [2].

We train two MLP models (with a hyperbolic tan activation function as it has better convergence properties than a logistic or sigmoid function) - one for the 2-class classification and one for the multi-class classification. Figure 15 shows the 2-class confusion matrix for a Multi-Layer Perceptron. Figure 18 shows the multi-class confusion matrix for a Multi-Layer Perceptron.

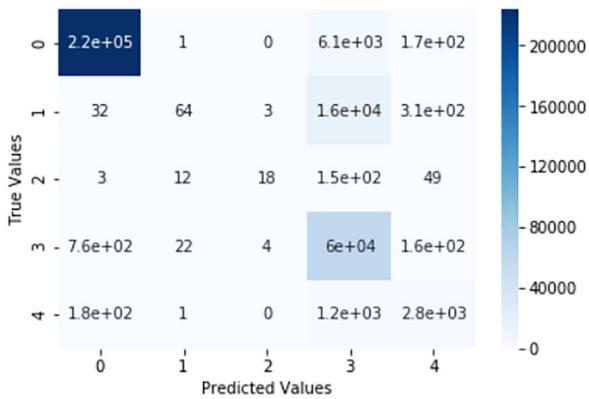
The overall accuracy, recall, precision and F1 score for the 2-class



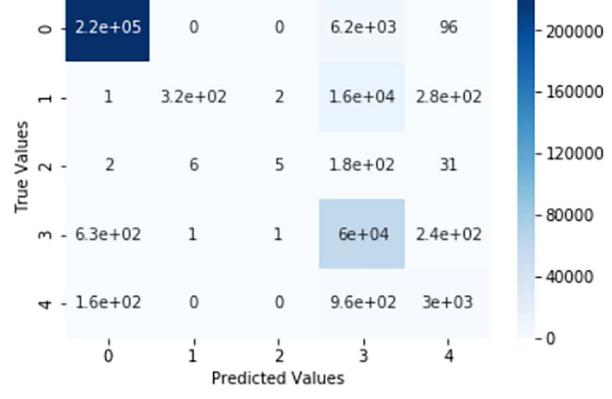
**Figure 17: Multi-Layer Perceptron Confusion Matrix - 2-class classification**



**Figure 19: Ensemble Model Confusion Matrix - 2-class classification**



**Figure 18: Multi-Layer Perceptron Confusion Matrix - Multi-class classification**



**Figure 20: Ensemble Model Confusion Matrix - Multi-class classification**

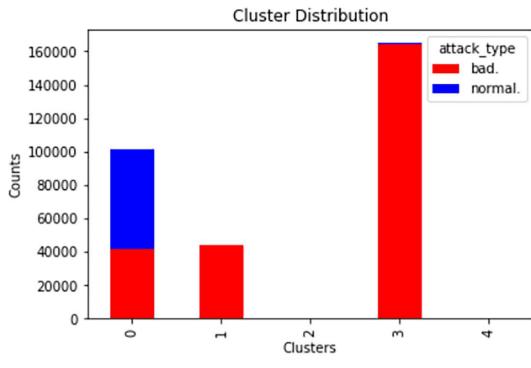
classification are as follows: 92.40, 94.68, 86.02 and 89.27 percent. The same for the multi-class classification are as follows: 91.98, 54.18, 77.53 and 53.90. We can observe that the accuracy of this model is similar to that of a random forest model for the 2-class classification. The recall and F1 score for multi-class classification is comparable to the random forest model.

**4.3.7 Ensemble Modeling.** Ensemble modeling deals with the combination of two or more machine learning models to generate a model with better accuracy. We have already observed that Random Forests have the highest accuracy for the 2-label classification whereas a polynomial SVM has better recall for the multi-label classification. Therefore, we try to get the best of both worlds by creating an ensemble of two Random Forest (with different rules for selection of the feature subset - square root of features and log of features) and one polynomial SVM model.

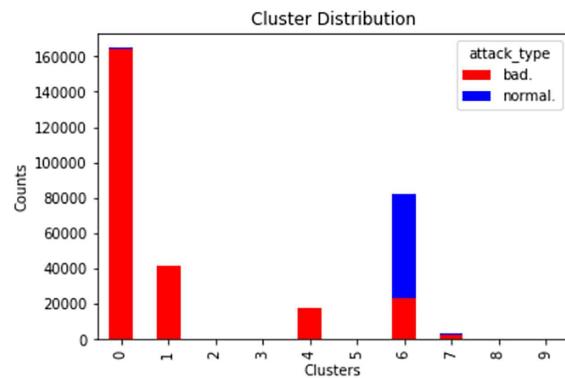
We train two ensemble models - one for the 2-class classification and one for the multi-class classification. Figure 19 shows the 2-class confusion matrix for an ensemble model. Figure 20 shows the

multi-class confusion matrix for an ensemble model. The overall accuracy, recall, precision and F1 score for the 2-class classification are as follows: 92.66, 95.25, 86.33 and 89.65 percent. The same for the multi-class classification are as follows: 92.26, 56.85, 84.77 and 55.41. We can observe that the accuracy and F1 score of this model is higher than all individual models for the 2-class classification. The recall and F1 score for multi-class classification is balanced between that of the random forest and the polynomial SVM but is higher than most individual models.

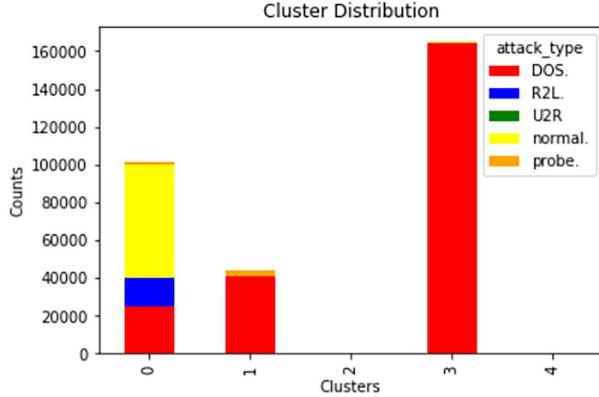
**4.3.8 Unsupervised Learning - Clustering.** Up till here, we observed and evaluated a variety of supervised learning models. As a result, we came to the conclusion that an ensemble of two good models often results in a better and more balanced result than individual models. In this section, we will examine how exploring the test data by means of a clustering algorithm (with no support from the training data) helps provide a good idea of the patterns within the data. The clustering algorithm we will use for this purpose is K-Means which is used to partition data into a given number of



**Figure 21: Chart for k-means clustering (clusters=5) - 2-class classification**



**Figure 23: Chart for k-means clustering (clusters=10) - 2-class classification**



**Figure 22: Chart for k-means clustering (clusters=5) - multi-class classification**

non-overlapping clusters based on a distance metric [4]. We train two k-means clustering models for both 2-class and multi-class classification - one for clusters=5 and the other for clusters=10 (for greater granularity).

Figure 21 shows the 2-class chart for k-means clustering with clusters=5 (some clusters not visible due to small size). We see that most of the clusters show one of the classes as a dominant proportion of the cluster. We can validate the same by comparing with the multi-class labels as well.

Figure 22 shows the multi-class chart for k-means clustering with clusters=5 (some clusters not visible due to small size).

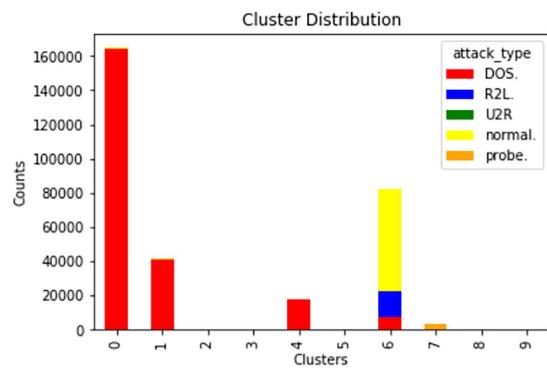
We also run the analysis for clusters=10, for greater granularity. Figure 23 shows the 2-class chart for k-means clustering with clusters=10 (some clusters not visible due to small size).

Figure 24 shows the multi-class chart for k-means clustering with clusters=10 (some clusters not visible due to small size).

We can observe the same trend here as well.

#### 4.4 Results

Among the supervised models, we observe that on comparison, some models perform better in terms of accuracy whereas some



**Figure 24: Chart for k-means clustering (clusters=10) - multi-class classification**

perform better in terms of recall. We also observe that most models find it easier to perform a 2-class classification (due to the high volume of attack labels in both the datasets as compared to normal labels), but face difficulties in identifying the individual classes (especially R2L and U2R which have a higher proportion in the test data compared to the training data). Overall, for the purpose of Dos/DDoS and intrusion detection, we see that most machine learning models give good results (KNN for example), and an ensemble of a random forest and polynomial SVM model gives the best accuracy among all.

When we venture into unsupervised learning we observe that clustering algorithms too can work well on network traffic data by creating clusters of traffic logs through pattern recognition. Though clustering does not provide us with exact labels, it can be useful in cases where we do not have any training or benchmark data, by giving us a fair idea of the direction in which to proceed.

#### 5 APACHE SPARK - USING PYSPARK

The volume of network traffic data generated is generally quite huge, and thus requires Big Data technologies to deal with it. Our demonstration was for a smaller subset of the actual dataset (which in itself consists of five million records). However, this larger dataset

too consists of logs only for seven weeks of monitoring. We can therefore imagine how voluminous the datasets would begin to get with constant monitoring of systems. In such cases, Big Data cloud technologies can come to the aid of analytics, and help create a sustainable system for such intrusion detection purposes.

Our analysis was carried out using Python on an individual system. But often for larger datasets, we need additional resources. The PySpark API, from Apache Spark (an open-source processing engine), can help us gain “access to the extremely high-performance data processing enabled by Spark’s Scala architecture - without the need to learn any Scala” [3]. The smallest building blocks of Spark are referred to as RDDs (Resilient Distributed Datasets) and these along with Spark’s DataFrame can act as useful alternatives to the *Pandas* data frames, in case of large datasets, where the distributed processing power of Spark can come into play [3].

We can install PySpark on a Windows machine using GOW (incorporates Linux commands in Windows like gzip, curl and tar) and Anaconda (an open-scale distribution containing Jupyter Notebook and other resources for Python). The package can be installed from the Apache Spark website, following which we perform *gzip* and *tar* operations on it. After adding the windows binary for Hadoop and modifying a few environment variables, you can launch Spark locally from Command Prompt. We have not used Spark for our analyses further as Python was able to handle the 10 percent datasets locally. However, PySpark can prove to be a great tool for analyzing data and creating models for larger datasets using a familiar and flexible language like Python. The presence of libraries like *mllib* in PySpark can offer us a wide variety of learning algorithms (similar to the *sklearn* library in Python).

## 6 CONCLUSION

The detection and prevention of DDoS attacks is a crucial problem for the safety and stability of networks. With the increasing use and dependence on technology and connectivity, this affects a huge cohort of people today. The data generated from day-to-day network traffic is huge and largely unstructured, but it can be captured and modified into an understandable structure, to be analyzed and used to generate efficient solutions. Through our analysis, we affirm the efficiency of machine learning technologies as tools for Big Data analytics and the use of open-source distributed processing systems as supports towards utilization of these tools. We observe that not only do supervised learning methods work well towards this objective, but unsupervised learning techniques such as clustering also provide us with helpful insights on pattern detection in the data. Therefore, Big Data technologies along with intelligent analytic solutions can help create new and improve existing defense systems to ensure security from such malicious attacks and intrusions.

## REFERENCES

- [1] Monowar H. Bhuyan, H. J. Kashyap, D. K. Bhattacharyya, and J. K. Kalita. 2014. Detecting distributed denial of service attacks: methods, tools and future directions. *Comput. J.* 57 (2014), 537–556. <https://doi.org/10.1093/comjnl/bxt031>
- [2] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- [3] IBM. 2016. *PySpark High-performance data processing without learning Scala*. IBM.
- [4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer, New York.
- <https://doi.org/10.1007/978-1-4614-7138-7>
- [5] Jupyter. 2017. The Jupyter Notebook. (2017).
- [6] KDDCup99. 1999. KDD Cup 1999 Data. (1999).
- [7] Andrew Kiggins and Jeffrey Lyons. 2016. *AWS Best Practices for DDoS Resiliency*. Amazon Web Services.
- [8] MITLincolnLaboratory. 1998. DARPA Intrusion Detection Evaluation. (1998).
- [9] scikit learn. 2017. scikit-learn - Machine Learning in Python. (2017).
- [10] Jessica Stone. 2017. The Best DDoS Protection Services. (July 2017).
- [11] Lea Toms. 2016. Closed for Business - the Impact of Denial of Service Attacks in the IoT. (Feb 2016).
- [12] Guoxing Zhang, Shengming Jiang, and Gang Wei. 2009. A prediction-based detection algorithm against distributed denial-of-service attacks. In *Proceedings of the International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, Vol. 1. Leipzig, Germany, 106fi?!110.

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in zhang05
(There was 1 warning)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.33.34] pdflatex report.tex
```

```
=====
latex report
```

```
[2017-12-16 09.33.44] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 18.3s.
```

```
=====
Compliance Report
```

```
=====
name: Rawat, Neha
hid: 224
paper1: Nov 3 17 100%
```

```
paper2: Nov 6 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 224 project (null) 5836 content.tex
wc 224 project (null) 5690 report.pdf
wc 224 project (null) 348 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
passed: False
```

```
find input{format/final}
```

---

```
6: \input{format/final}
```

```
passed: True
```

```
floats
```

---

```
29: \begin{figure}
30: \includegraphics[width=1.0\columnwidth]{images/DDoS.PNG}
32: \label{F:ddos}
34: Figure \ref{F:ddos} shows how a Distributed Denial-of-Service
   attack occurs.\\
40: \begin{figure}
```

```

41: \includegraphics[width=1.0\columnwidth]{images/OSI.PNG}
43: \label{F:osi}
45: Figure \ref{F:osi} shows an Open Systems Interconnection Model
   with the layers highlighted where DDoS attacks are most common.\\
53: \begin{figure}
54: \includegraphics[width=1.0\columnwidth]{images/dossteps.PNG}
56: \label{F:doss}
58: Figure \ref{F:doss} shows the steps of a Denial-of-Service attack
   execution.\\
62: \begin{figure}
63: \includegraphics[width=1.0\columnwidth]{images/dosattackstat.PNG}
65: \label{F:dosstat}
67: Figure \ref{F:dosstat} shows the percentage of different Denial-
   of-Service attacks in 2011 by type.\\
93: \begin{figure}
94: \includegraphics[width=1.0\columnwidth]{images/attack_type.PNG}
96: \label{F:att}
98: Figure \ref{F:att} shows the Attack Type distribution in the
   training and test datasets\\
101: \begin{figure}
102: \includegraphics[width=1.0\columnwidth]{images/pairplot.png}
104: \label{F:pair}
106: Figure \ref{F:pair} shows the pair plot between the first 15
   variables in the training dataset. We observe that the data seems
   to be skewed, indicating the need for standardizing the features.
   Also, there do not seem to be a lot of correlated variables in
   the dataset.\\
115: \begin{figure}
116: \includegraphics[width=1.0\columnwidth]{images/logreg2.PNG}
118: \label{F:logreg2}
120: Figure \ref{F:logreg2} shows the 2-class confusion matrix for
   logistic regression.
121: \begin{figure}
122: \includegraphics[width=1.0\columnwidth]{images/logregall.PNG}
124: \label{F:logregall}
126: Figure \ref{F:logregall} shows the multi-class confusion matrix
   for logistic regression.\\
134: \begin{figure}
135: \includegraphics[width=1.0\columnwidth]{images/knn2.PNG}
137: \label{F:knn2}
139: Figure \ref{F:knn2} shows the 2-class confusion matrix for KNN.
140: \begin{figure}
141: \includegraphics[width=1.0\columnwidth]{images/knnaall.PNG}
143: \label{F:knnaall}
145: Figure \ref{F:knnaall} shows the multi-class confusion matrix for
   KNN.\\

```

```
152: \begin{figure}
153: \includegraphics[width=1.0\columnwidth]{images/svm2.PNG}
155: \label{F:linsvm2}
157: Figure \ref{F:linsvm2} shows the 2-class confusion matrix for
linear SVM.
158: \begin{figure}
159: \includegraphics[width=1.0\columnwidth]{images/svsmall.PNG}
161: \label{F:linsvmall}
163: Figure \ref{F:linsvmall} shows the multi-class confusion matrix
for linear SVM.\\
168: \begin{figure}
169: \includegraphics[width=1.0\columnwidth]{images/svmpoly2.PNG}
171: \label{F:polysvm2}
173: Figure \ref{F:polysvm2} shows the 2-class confusion matrix for
polynomial SVM.
174: \begin{figure}
175: \includegraphics[width=1.0\columnwidth]{images/svmpolyall.PNG}
177: \label{F:polysvmall}
179: Figure \ref{F:polysvmall} shows the multi-class confusion matrix
for polynomial SVM.\\
186: \begin{figure}
187: \includegraphics[width=1.0\columnwidth]{images/rf2.PNG}
189: \label{F:rf2}
191: Figure \ref{F:rf2} shows the 2-class confusion matrix for a
Random Forest.
192: \begin{figure}
193: \includegraphics[width=1.0\columnwidth]{images/rfall.PNG}
195: \label{F:rfall}
197: Figure \ref{F:rfall} shows the multi-class confusion matrix for a
Random Forest.\\
204: \begin{figure}
205: \includegraphics[width=1.0\columnwidth]{images/nn2.PNG}
207: \label{F:nn2}
209: Figure \ref{F:nn2} shows the 2-class confusion matrix for a
Multi-Layer Perceptron.
210: \begin{figure}
211: \includegraphics[width=1.0\columnwidth]{images/nnall.PNG}
213: \label{F:nnall}
215: Figure \ref{F:nnall} shows the multi-class confusion matrix for a
Multi-Layer Perceptron.\\
222: \begin{figure}
223: \includegraphics[width=1.0\columnwidth]{images/ensemble2.PNG}
225: \label{F:en2}
227: Figure \ref{F:en2} shows the 2-class confusion matrix for an
ensemble model.
228: \begin{figure}
```

```

229: \includegraphics[width=1.0\columnwidth]{images/ensembleall.PNG}
231: \label{F:enall}
233: Figure \ref{F:enall} shows the multi-class confusion matrix for
    an ensemble model.\\
240: \begin{figure}
241: \includegraphics[width=1.0\columnwidth]{images/cluster52graph.PNG}
    }
243: \label{F:cg52}
245: Figure \ref{F:cg52} shows the 2-class chart for k-means
    clustering with clusters=5 (some clusters not visible due to
    small size).\\
247: \begin{figure}
248: \includegraphics[width=1.0\columnwidth]{images/cluster5allgraph.P
    NG}
250: \label{F:cg5all}
252: Figure \ref{F:cg5all} shows the multi-class chart for k-means
    clustering with clusters=5 (some clusters not visible due to
    small size).\\
254: \begin{figure}
255: \includegraphics[width=1.0\columnwidth]{images/cluster102graph.PN
    G}
257: \label{F:cg102}
259: Figure \ref{F:cg102} shows the 2-class chart for k-means
    clustering with clusters=10 (some clusters not visible due to
    small size).\\
260: \begin{figure}
261: \includegraphics[width=1.0\columnwidth]{images/cluster10allgraph.
    PNG}
263: \label{F:cg10all}
265: Figure \ref{F:cg10all} shows the multi-class chart for k-means
    clustering with clusters=10 (some clusters not visible due to
    small size).\\

```

figures 24

tables 0

includegraphics 24

labels 24

refs 24

floats 24

True : ref check passed: (refs >= figures + tables)

True : label check passed: (refs >= figures + tables)

True : include graphics passed: (figures >= includegraphics)

True : check if all figures are referred to: (refs >= labels)

Label/ref check

passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

---

below\_check

---

WARNING: code and above may be used improperly

52: When launching a DDoS attack, the attacker goes through some steps common to both types of architectures \cite{monowar01}. First, the attacker tries to identify vulnerable systems that can be used as agents. The resources of these systems are used to generate a powerful attack stream. Next, the attacker plants the handler software code in the compromised system and ensures steps to prevent the code from being detected. These compromised systems are often referred to as {\em zombies}. Sometimes, the attacker creates several intermediate layers between the {\em zombies} and the victim to hinder traceability. Thirdly, the attacker communicates with the handler codes placed via protocols like TCP or UDP, and decides the scheduling of the attacks. Post the complete setup, the attacker launches the attack on the victim's machine or server and renders it unusable \cite{monowar01}. In an IRC-based architecture, most of the above steps remain same, but an IRC-channel is used for communication purposes. This helps the attacker as even if one {\em zombie} or {\em bot} is discovered, the identities of the others is still hidden, as IRC-channels are difficult to detect \cite{monowar01}.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in zhang05
(There was 1 warning)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 8217

---

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True

This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate

Missing character: ""

Missing character: ""

Missing character: ""

Typesetting of "report.tex" completed in 18.8s.

```
=====
Compliance Report
=====
```

```
name: Rawat, Neha
hid: 224
paper1: Nov 3 17 100%
paper2: Nov 6 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 224 project (null) 5836 content.tex
wc 224 project (null) 5726 report.pdf
wc 224 project (null) 348 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
passed: False
```

```
find input{format/final}
```

---

```
6: \input{format/final}
```

```
passed: True
```

floats

---

29: \begin{figure}  
30: \includegraphics[width=1.0\columnwidth]{images/DDoS.PNG}  
32: \label{F:ddos}  
34: Figure \ref{F:ddos} shows how a Distributed Denial-of-Service  
attack occurs.\\"  
40: \begin{figure}  
41: \includegraphics[width=1.0\columnwidth]{images/OSI.PNG}  
43: \label{F:osi}  
45: Figure \ref{F:osi} shows an Open Systems Interconnection Model  
with the layers highlighted where DDoS attacks are most common.\\"  
53: \begin{figure}  
54: \includegraphics[width=1.0\columnwidth]{images/dossteps.PNG}  
56: \label{F:doss}  
58: Figure \ref{F:doss} shows the steps of a Denial-of-Service attack  
execution.\\"  
62: \begin{figure}  
63: \includegraphics[width=1.0\columnwidth]{images/dosattackstat.PNG}  
65: \label{F:dosstat}  
67: Figure \ref{F:dosstat} shows the percentage of different Denial-  
of-Service attacks in 2011 by type.\\"  
93: \begin{figure}  
94: \includegraphics[width=1.0\columnwidth]{images/attack\_type.PNG}  
96: \label{F:att}  
98: Figure \ref{F:att} shows the Attack Type distribution in the  
training and test datasets\"  
101: \begin{figure}  
102: \includegraphics[width=1.0\columnwidth]{images/pairplot.png}  
104: \label{F:pair}  
106: Figure \ref{F:pair} shows the pair plot between the first 15  
variables in the training dataset. We observe that the data seems  
to be skewed, indicating the need for standardizing the features.  
Also, there do not seem to be a lot of correlated variables in  
the dataset.\\"  
115: \begin{figure}  
116: \includegraphics[width=1.0\columnwidth]{images/logreg2.PNG}  
118: \label{F:logreg2}  
120: Figure \ref{F:logreg2} shows the 2-class confusion matrix for  
logistic regression.  
121: \begin{figure}  
122: \includegraphics[width=1.0\columnwidth]{images/logregall.PNG}  
124: \label{F:logregall}  
126: Figure \ref{F:logregall} shows the multi-class confusion matrix  
for logistic regression.\\"

```

134: \begin{figure}
135: \includegraphics[width=1.0\columnwidth]{images/knn2.PNG}
137: \label{F:knn2}
139: Figure \ref{F:knn2} shows the 2-class confusion matrix for KNN.
140: \begin{figure}
141: \includegraphics[width=1.0\columnwidth]{images/knnall.PNG}
143: \label{F:knnall}
145: Figure \ref{F:knnall} shows the multi-class confusion matrix for
KNN.\\
152: \begin{figure}
153: \includegraphics[width=1.0\columnwidth]{images/svm2.PNG}
155: \label{F:linsvm2}
157: Figure \ref{F:linsvm2} shows the 2-class confusion matrix for
linear SVM.
158: \begin{figure}
159: \includegraphics[width=1.0\columnwidth]{images/svsmall.PNG}
161: \label{F:linsvsmall}
163: Figure \ref{F:linsvsmall} shows the multi-class confusion matrix
for linear SVM.\\
168: \begin{figure}
169: \includegraphics[width=1.0\columnwidth]{images/svmpoly2.PNG}
171: \label{F:polysvm2}
173: Figure \ref{F:polysvm2} shows the 2-class confusion matrix for
polynomial SVM.
174: \begin{figure}
175: \includegraphics[width=1.0\columnwidth]{images/svmpolyall.PNG}
177: \label{F:polysvsmall}
179: Figure \ref{F:polysvsmall} shows the multi-class confusion matrix
for polynomial SVM.\\
186: \begin{figure}
187: \includegraphics[width=1.0\columnwidth]{images/rf2.PNG}
189: \label{F:rf2}
191: Figure \ref{F:rf2} shows the 2-class confusion matrix for a
Random Forest.
192: \begin{figure}
193: \includegraphics[width=1.0\columnwidth]{images/rfall.PNG}
195: \label{F:rfall}
197: Figure \ref{F:rfall} shows the multi-class confusion matrix for a
Random Forest.\\
204: \begin{figure}
205: \includegraphics[width=1.0\columnwidth]{images/nn2.PNG}
207: \label{F:nn2}
209: Figure \ref{F:nn2} shows the 2-class confusion matrix for a
Multi-Layer Perceptron.
210: \begin{figure}
211: \includegraphics[width=1.0\columnwidth]{images/nnall.PNG}

```

```

213: \label{F:nnall}
215: Figure \ref{F:nnall} shows the multi-class confusion matrix for a
     Multi-Layer Perceptron.\\
222: \begin{figure}
223: \includegraphics[width=1.0\columnwidth]{images/ensemble2.PNG}
225: \label{F:en2}
227: Figure \ref{F:en2} shows the 2-class confusion matrix for an
     ensemble model.
228: \begin{figure}
229: \includegraphics[width=1.0\columnwidth]{images/ensembleall.PNG}
231: \label{F:enall}
233: Figure \ref{F:enall} shows the multi-class confusion matrix for
     an ensemble model.\\
240: \begin{figure}
241: \includegraphics[width=1.0\columnwidth]{images/cluster52graph.PNG}
    }
243: \label{F:cg52}
245: Figure \ref{F:cg52} shows the 2-class chart for k-means
     clustering with clusters=5 (some clusters not visible due to
     small size).\\
247: \begin{figure}
248: \includegraphics[width=1.0\columnwidth]{images/cluster5allgraph.PNG}
250: \label{F:cg5all}
252: Figure \ref{F:cg5all} shows the multi-class chart for k-means
     clustering with clusters=5 (some clusters not visible due to
     small size).\\
254: \begin{figure}
255: \includegraphics[width=1.0\columnwidth]{images/cluster102graph.PNG}
257: \label{F:cg102}
259: Figure \ref{F:cg102} shows the 2-class chart for k-means
     clustering with clusters=10 (some clusters not visible due to
     small size).\\
260: \begin{figure}
261: \includegraphics[width=1.0\columnwidth]{images/cluster10allgraph.PNG}
263: \label{F:cg10all}
265: Figure \ref{F:cg10all} shows the multi-class chart for k-means
     clustering with clusters=10 (some clusters not visible due to
     small size).\\

```

figures 24

tables 0

includegraphics 24

labels 24

```
refs 24
floats 24
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
WARNING: code and above may be used improperly
```

52: When launching a DDoS attack, the attacker goes through some steps common to both types of architectures \cite{monowar01}. First, the attacker tries to identify vulnerable systems that can be used as agents. The resources of these systems are used to generate a powerful attack stream. Next, the attacker plants the handler software code in the compromised system and ensures steps to prevent the code from being detected. These compromised systems are often referred to as {\em zombies}. Sometimes, the attacker creates several intermediate layers between the {\em zombies} and the victim to hinder traceability. Thirdly, the attacker communicates with the handler codes placed via protocols like TCP or UDP, and decides the scheduling of the attacks. Post the complete setup, the attacker launches the attack on the victim's machine or server and renders it unusable \cite{monowar01}. In an IRC-based architecture, most of the above steps remain same, but an IRC-channel is used for communication purposes. This helps the attacker as even if one {\em zombie} or {\em bot} is discovered, the identities of the others is still hidden, as IRC-channels are difficult to detect \cite{monowar01}.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in zhang05
(There was 1 warning)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

non ascii found 8217

---

The following tests are optional

---

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

```
find cite {
```

---

```
passed: True
```

# Predictive Model For English Premier League Games

Josh Lipe-Melton  
Indiana University  
4400 E Sheffield Dr  
Bloomington, Indiana 47408  
jlipemel@umail.iu.edu

## ABSTRACT

We discuss a model for predicting the outcome of soccer matches based on the previous matches played by each team. The model we produce is based on a model discussed in a previous paper, which claims to predict match results extremely accurately based on the previous two meetings between the teams and the previous five matches of each of the teams. While the model we based the project on used a genetic tuning algorithm combined with a neural network, this was too complicated for our purposes. Instead, we first attempted to use a multivariate regression model. This model uses matrix algebra to generate coefficients based on the sample data given to it. These coefficients are then what is used to make predictions. The next model was a basic neural network model. This model creates different layers of perceptrons, which can solve more complicated problems than a single perceptron. Both of these models make use of python's sklearn package, and the code for our model is in the project.ipynb file. We evaluate both models and compare the predictive accuracy of each using a number of metrics. Potential uses for this type of model could include setting gambling lines or for the evaluation of the importance of certain games relative to one another. Furthermore, the analysis of the data could be used on any sets of data which are in the dataframe format. The neural network, multivariate regression model, and evaluation code is very flexible and could easily be used for other purposes.

## KEYWORDS

HID105, I523, sports, analytics, predictive, neural network

## 1 INTRODUCTION

Prediction of sporting events is an extremely difficult problem due to the enormous number of factors involved and the unpredictability of those factors. While a lot of data about sports is generated, it is still extremely difficult to create models which account for every factor involved. In the model we attempt to imitate, a match's result is hypothesized to be predictable based on the last five matches each team has played and the last two matches between the teams. The model we imitated created a three layer neural network using these features, and was initiated with weights created by a genetic tuning algorithm. The model we use loads data on the English Premier League, England's highest league and arguably the best league in the world. Using this data, we create features for each match based on the previous games played by those teams. In this way, we attempt to circumvent far more complicated methods of analyzing sports such as per possession models or spatial recognition and player tracking software. Common thought among soccer players indicates that a team's 'form', or how well they have played in their

last five matches, is a significant indicator of how well a team will do in their next match. Similarly, it seems to follow common sense that if team 1 has beaten team 2 the previous two times they have played, that team 1 is likely to beat team 2 again the next time they play.

There are many supporting factors to the assumption that a team that has beaten another repeatedly will do so again. For one, teams in the premier league with more money consistently do significantly better than those with less money. According to Gerhards, "success in national football championships is highly predictable. The market value of a team is by far the most important single predictor". [3] In comparison to variables such as diversity of a team or the amount of turnover in team personnel, market value was far more positively correlated to success. Furthermore, those teams that find success sell more jerseys and are featured on television more, thus generating even more wealth and ensuring that the wealth generated by soccer stays with the wealthier teams. These facts seem to support the statement that plugging the form, or last five results of each team, and the previous results between the two teams, into some model, we could expect to predict with some consistency the results of soccer matches. Market value implies consistent success of some teams over others, which would seem to indicate that the previous two meetings between two teams should consistently correlate to results. Given previous results, some of the unpredictability of sporting events in general is taken away. While random chance should still certainly be accounted for, we expect significant correlation between previous match results and future match results.

Although numerous other statistics are available that describe English Premier League soccer in more detail, we chose not to incorporate any other variables in our models. In general, the models discussed in paper 2 which performed well used fewer variables. The models that used a lot of different kinds of data were less effective. Therefore, the project model used only match results as data. In order to create models for predictive analytics, we used sklearn's python packages. The first is the multivariate linear regression model, which takes any number of variables and weights them according to their correlation to the true results. This model produces a continuous range of predictions. We also created a similar linear regression model using less features. Lastly, we created a neural network using the sklearn neural network package. This package takes an array of inputs, in this case match results, and produces layers of perceptrons, or 'neurons'. By combining multiple 'neurons' making use of stochastic gradient descent, more complicated problems and models can be represented than when using just one. Furthermore, "The use of SGD In the neural network setting is motivated by the high cost of running back propagation over the full training set. SGD can overcome this cost and still lead

to fast convergence.”[8] Although sklearn offers various additional variables for tuning implementations of its neural network package, we chose not to use them in our code. This was due to a lack of experience tuning neural networks.

## 2 LEARNING MODELS EXAMINED FROM PAPER 2

In order to come up with our model, we examined models in paper 2 in order to develop a strategy to solve the complicated problem of predicting soccer games. We evaluated effectiveness and ease of implementation. Furthermore, we evaluated the different forms of data and models used in order to learn and implement the best parts of each model in our own model. This section is largely excerpts from paper 2.

### 2.1 Expected Goals Model

Arguably the most common method of predicting the results of soccer games is to create a prediction of the number of goals scored by each team. The result of subtracting these two numbers gives not only a prediction of which team will win, but an inherent level of confidence proportional to the difference of each predicted number of goals [1]. This model creates an “expected goals value” by predicting the number of shots and assigning each of these shots a value. These values are based on attributes such as angle from the goal, distance to the goal, body part used to take the shot, what type of approach was used to obtain the shot (dribble, short pass, long pass, etc.), and even the relevant FIFA video game ratings of the player taking the shot. Each value represents the predicted likelihood of scoring, with 0 being an impossible shot and 1 being a sure goal. By summing these values and incorporating the FIFA rating of the opposing goalkeeper, an expected goals value for a team is obtained. This model is able to predict the number of goals scored by each team about 20% of the time. The correct result of the match was found about 56% of the time [1]. While the data was extremely specific, the general assumption that a team’s goals in a given match correlate to the quality of the shots the team gets plus the quality of the striker was extremely ineffective. [5]

### 2.2 Bivariate Expected Goals Model

We drew some inspiration from this model discussed in paper 2: A flaw with the previous example of an expected goals model is that it accounted only for the attack team’s ability in its goal predictions. Apart from the ability of the goalkeeper, there is no accounting for the defensive ability of an opponent in prediction of expected goals. In a different model, defensive ability and attacking ability are both incorporated. The authors of this method created their model based on the idea that the goals scored two competing soccer teams are negatively correlated with one another. By using a bivariate Poisson model for soccer data, the authors created predictions for the number of goals scored by each team in a given match, and therefore the results of each game[4]. The covariates used in the bivariate Poisson regression model include: GDP per capita, population, home advantage, bookmaker’s odds, market value, number of Champion’s League players, number of club teammates, and the age of the coach. By running 1,000,000 simulations on the European Championships in 2016, predictions for each match were created,

along with odds for each team to reach each round of the tournament. The odds of the model outperformed bookmakers’ odds 42.22% to 39.23% in predictive accuracy[4]. The authors used their model in placing equal bets on every bet in the tournament with the service that provided the most favorable odds to the outcome predicted by their model. In doing so, they obtained a return of 30.28% after the tournament. The authors concluded that the scores of two soccer teams are indeed negatively correlated and that this is a sound notion to base a predictive model on [4]. [5]

From that model, we gained insights into how to make our model. First of all, the authors of that model gave an example that used a relatively simple to implement bivariate poisson model. Secondly, the authors of the model concluded that two team’s goals were strongly negatively correlated. It is therefore important to take both teams into equal consideration. Finally, the authors of this model classified their results into simply home team win, home team loss, or draw, which provided an easy and effective way to evaluate the model. Our model is similarly able to predict a home win, home loss, or home tie, and the results are evaluated in this way as well.

### 2.3 NCAA Analysis

In college basketball, the committee that decides who gets into the NCAA tournament makes use of a ranking system called Ratings Percentage Index, or RPI. RPI weights .25 of a team’s ranking on their win percentage, .5 on their opponents’ win percentage, and .25 on their opponents’ opponents’ win percentage. [9] This system is designed to encourage teams to schedule difficult opponents, as a large portion of the rankings is based on strength of schedule. This formula has significant influence on where teams are ranked. Unfortunately, “the RPI lacks theoretical justification from a statistical standpoint.” [9] In general, it is believed that the model places too much emphasis on strength of schedule and not enough on performance. Attempts to utilize an improved version of this model have made an impact on seeding in college soccer and baseball as well. In these sports, wins are weighted to give more ranking points to an away win than a home win.[9] These types of alterations, however, do not address the fact that 75% of this ranking comes from a team’s strength of schedule. This type of bias favors teams that are in strong conferences, even if they have poor records in their conference. [5]

### 2.4 Per Possession Analysis

A proposed alternative to RPI is to use a “per possession model,” or a model that predicts outcomes using statistics that are used in the context of efficiency with possessions. For example, offensive efficiency is found by dividing points scored by possessions and defensive efficiency is found by dividing points allowed by possessions [10]. These statistics are then used to calculate an offensive efficiency adjusted by the perceived strength of the opponent. Adjusted offensive efficiency, for example, is calculated by multiplying offensive efficiency by the average national offensive efficiency then dividing this number by the adjusted defensive efficiency of an opponent [10]. By combining these adjusted efficiencies with other factors such as home court advantage, the authors made several models which created an estimation for “win probability,”

which can in turn be used to predict individual match outcomes or create a ranking system. By using win probability, the study we examine created models based on decision trees, rule learners, artificial neural networks, naive Bayes, and ensemble learners. The neural network and naive Bayes models were the most effective models, both predicting outcomes with about 72% accuracy[10]. A surprising observation from the authors is that simpler models tend to work better than more complicated ones. Similarly, attempting to incorporate more features into the models tended to decrease predictive accuracy. The authors believe that there is a "glass ceiling" when it comes to accuracy predicting sporting events of around 74% [10]. Each of these models is unable to predict any individual season at a rate greater than 74%[10]. [5]

## 2.5 Fuzzy Neural Network Model

In a paper 2, we discussed a method of prediction solely uses past results to predict future results. This model was extremely accurate and a under strong consideration for a model to base our project on. This section is an excerpt from [5]: In this method, a predictive model is based on the intuitive proposition that if team 1 has won their previous few games, team 2 has lost their previous few games, and team 1 has beaten team 2 the last two times they have played, team 1 will beat team 2 [6]. The model proposed in this article assigns a value in the range [-5, 5] to the last five games played by each team as well as the last two games played between the two teams. The higher the number, the bigger the win. The lower the number, the bigger the loss. The predicted result of a game is a function of these numbers. Through a combination of a fuzzy logic table and a neural network algorithm, a result is predicted. First, the authors created a table with every possible value of  $x_1 \times 12$ . Each of these combinations was then associated with a predicted result and a weight in the interval [0, 1] that indicated the confidence in the predicted result. These initial confidence intervals were then tuned. The predicted result is drawn from the range [Big loss (BL), Small loss (SL), Draw (D), Small win (SW), Big win (BW)] [6]. Using a sample size of 1056 matches, the network assigned weights to the nodes in the neural network. The trained model was applied to 350 results from other seasons and was correct when predicting a big loss 91.4 percent of the time, a small loss 83.3 percent of the time, a draw 87 percent of the time, a small win 84 percent of the time, and a big win 94.6 percent of the time [6]. The authors do cite flaws that come from not considering factors such as injured or suspended players, refereeing, or weather conditions [6]. [5]

Furthermore, this method's already impressive predictive accuracy could also be improved by taking into account strength of schedule, as a team that has narrowly won its last five games against very weak opponents would be favored against a team that has narrowly lost against very strong opponents. The machine learning techniques implemented in this study could have been improved by incorporating opponents' results into the model, giving more weight to wins against good teams. [5] It would also be interesting to see whether using a continuous model would decrease the accuracy of predictions or give similar accuracy with more specificity than the fuzzy logic model. It seems possible that using the fuzzy logic model provides a neural network with more occurrences of samples that are similar to each other due to grouping results

together, thereby providing a better prediction. In a continuous model, the features may be too varied for a neural network to pick up on without a greatly increased sample size.

After examining these models, we chose to create a multivariate linear regression model and a neural network. These models would use match results. The predictions would be continuously distributed, and would indicate the degree to which the home team would be expected to win, lose, or draw by. Python was chosen to implement these solutions due to the ease of use in machine learning and data analytics applications, as well as being the default language for this course. The Sklearn package was chosen due to the ease of implementation. Instead of having to construct a neural network or linear regression model from scratch, the implementation was straightforward and contained many ways to customize the models they provided. Our predictions will be evaluated based on percent of correct match results predicted, matches correctly predicted within 1 goal, matches correctly predicted to within half a goal, and mean squared error.

## 3 PROJECT MODEL

In order to select the parameters for our model, we examined several models in paper 2. The most common type of predictive model for other sports was a per possession model. This model attempts to gauge the number of possessions each team will get, then gauge how efficient each team will be with their possessions. By multiplying the number of predicted possessions by the projected efficiency of the team, a prediction for the number of points scored by that team occurs. This means that a match prediction would be the difference of the predicted goals for each team. In soccer, this could be done with time of possession, a commonly tracked statistic. A model could for example predict that for each minute a team is expected to possess the ball, they are expected to score a certain number of goals. By incorporating the opposing team's predicted minutes of possession and predicted goals conceded, a prediction of goals scored and conceded could be obtained. However, after considering this type of model, we decided it had too many flaws to be implemented. Firstly, the model would not use a simple, readily available set of data. Secondly, the model would have to incorporate a greatly varying set of data. In paper 2, we concluded that "simple inputs, especially those involving neural networks, provide the greatest accuracy in predicting the outcome of sporting events." [5] Therefore, we decided to reject models that had several forms of data and stick to only match data. It was also important to figure out what kind of data to use. In paper 2, models used features such as FIFA ratings, possession statistics, match results, and expected goals. We decided to move forward using match results due to the simplicity of that variable, as well as the abundance and ease of access for that data in a number of leagues. Ultimately, we chose to focus on just one league in order to attempt to keep the data consistent and to try to make predictions based on the highest level soccer possible. It could be an interesting topic to compare models' effectiveness in evaluating other leagues, but we chose to use match data from just the English Premier League.

In general, we attempted to imitate the model discussed in the section 'Fuzzy Neural Network Model'. We used match data from <http://www.football-data.co.uk> [2]. The data includes numerous

statistics about English Premier League soccer matches dating back to 1993, including the team names and goals scored by each team, which were the data we were interested in. In order to load the data from the .csv files included in the website into a usable format, we used panda's read csv function. Three years of the data was in an unreadable format for the csv loader and was skipped in the analysis, which included data from 2001-2003. We then narrowed the data down to the names of each team involved in each match and the number of goals scored by each team. We then created a function last5 to determine the last five matches the last team played, which returned the number of goals scored by the team minus the number of goals that had been scored on them over the course of those five games. This function partitioned off the dataframe with all results to include just the results which had occurred before the match in question. Next, we used a boolean indexer in the dataframe to determine which rows contained the team name in either the home team or away team column. Each entry was added to a different list. That list was then negatively indexed to find the last five entries in the list, and each item was summed. This sum was what the function returned. We also created a function prevMeetings to determine the results of the last two times the teams played. This function took the data and the index of the match to be observed. Next, the names of the teams involved were found. Next, the data was slimmed to only include match results from the past. Next, we used a boolean indexer in the dataframe of past matches to determine which rows contained the home team name in either the home team column or the away team column, and the away team name in either the home team column or the away team column. We negatively indexed the dataframe produced to include only the last two results. Each of these results was summed. In each of these functions, we used goal difference to represent each result. Goal difference was found by subtracting the number of goals allowed in a match by the home team from the number of goals scored in a match by the home team. We kept individual match results within the range [-4, 4] in order to prevent very large wins or losses from having too much influence on the statistics. A 7-0 win, therefore, counted the same as a 4-0 win, and a 7-0 win followed by five losses couldn't result in a positive goal difference for the team. Next, we created a function sampler, which used prevMeetings and last5 to turn each row in the data into an array of the features about each match that we wanted. We referred to the last5 of the home team as 'z1', the last5 of the away team as 'z2', and the previous two meetings of the teams as 'z3' and 'z4' respectively in order to more consistently imitate the model discussed in the Fuzzy Neural Network Model. Finally, we created a function that found the true results of each match by iterating through every match result and subtracting the away team's goals scored from the home team's goals scored, thus representing the "true" prediction. This model has the benefit of inherently giving value to the home team due to z1 always being the home team and z2 always being the away team. When two teams did not have previous results, z3 and z4 were entered as 0 and 0 so as not to affect the prediction either way. Because of this, we trimmed the first 250 results out of the data so as not to have too many z3 and z4 data points equal to 0. When a team did not have five previous matches in the data, last5 returned a -5, as this typically indicates a team recently promoted to the league and therefore the team would not

be expected to find much success. These functions slightly differed from the more complicated model we were imitating, as this used the last five results from each team as individual features in the first input layer of a neural network and the previous two meetings as input nodes in the second layer of a neural network. Furthermore, the model we were imitating used fuzzy logic to model the problem as a classification problem, using big loss, small loss, draw, big win, and small win as the classifications. Our model, however, attempts to create a continuous solution. In order to do so, we take the results of sampler as our sample data used for prediction and the output of another function, results, as our true data. Sklearn's multivariate linear regression model uses matrix algebra in order to create coefficients for each variable in the sample X. Each coefficient indicates the strength of the correlation between the variable in X and the actual result. Therefore, the coefficient is the weight which the variable is multiplied by when using the model to predict. Using sklearn's model, we fitted the sample data to the results and created an array of predictions, which we then compared to the true results. We also used sklearn's neuralnetwork package to create a MLP Regressor neural network with the hidden layer sizes attribute set to 50, which we found to produce the smallest mean squared error. In order to reduce the bias towards z1 and z2, which are typically bigger in absolute value than z3 and z4, the data was scaled during preprocessing for this model. Next, the neural network fit the scaled data to the true results. Finally, evaluation of the model was done based on mean squared error between predictions and true results and percentages of correct predictions or predictions that were within a certain range of the true result. In our model, a correct prediction was classified as simply predicting within the same category as the result, with the categories being less than -.5, greater than .5, or in between .5 and -.5 goal difference, each representing a draw, home win, or home loss respectively. We also tested whether each prediction was within .5 or 1 of the true result. These predictions represent the expected goal difference of the home team, meaning the predicted value of the home team's goals scored minus the away team's goals scored. After the prediction sets for each model were created, we changed the mean of the data to fit the mean goal difference of the results, which was just over 0.4. We also changed the standard deviation to match the standard deviation of the results.

## 4 EVALUATION

### 4.1 Efficiency

Extraction of the data was relatively fast. The retrieveEPL function extracts 7832 rows of a dataframe relatively quickly. The last5 function is slow due to running 7832 times and checking a large portion of the dataframe for previous results each time. The sampler function is by far the slowest to run due to the large number of comparisons and indexes it has to do. Running last5 and prevMeetings on each row of the dataframe is less than quadratic time, but still ends up being extremely costly computationally. The results function runs in linear time, which is optimal. Finally, the models are fitted to the data extremely quickly as well.

## 4.2 Prediction Performance

The performance of the predictions was far less effective than the fuzzy neural network model we attempted to replicate. This was to be expected, however, as our model was simply a more basic version of the other model. In order to evaluate our model, we measure the mean squared error between the actual results and the predicted results of both the multivariate linear regression model and the neural network. Each had an almost identical mean squared error, with the neural network scoring 2.851 and the linear regression model scoring 2.868. Each model also had nearly identical proportions of games predicted within .5 of the correct result (.264 for the neural network and .263 for the multivariate linear regression model), games predicted within 1 of the correct result (.49 for both models), draws correctly predicted (.147 for the neural network and .144 for the multivariate linear regression model), home wins correctly predicted (.249 for the neural network and .255 for the multivariate linear regression model), home losses correctly predicted (.020 for the neural network and .019 for the multivariate linear regression model), and percent total correct predictions (.417 for the neural network and .419 for the multivariate linear regression model). The effect of z1 and z2 was as would be expected: z1 was a positive coefficient for the regressino model and z2 was negative. This indicates that the home team was favored, which was reflected in the mean value in both the prediction and true results sets being about 0.4. It was interesting to note, however, that the z3 and z4 features had almost no effect on the accuracy of the predictions. The linear regression model's coefficient for these two features was almost 0. This is extremely counter intuitive and could be a result of inserting 0 for z3 and z4 at times where no previous results could be found, although it seems that these should still be significant features for prediction. Both models' prediction sets' standard deviation was far lower than the actual results, which could be a significant factor in the mean squared error. After hypothesizing that this may have increased the number of home losses predicted correctly, we tried increasing the standard deviation. After subtracting the mean of each array from each element in the array, then multiplying each value by 1.76, the standard deviation of the true results, and adding the mean back to each element, the percent of correct results predicted increased to 45.00 and 45.50 percent for the multivariate linear regression model and neural network model respectively. Doing this, however, decreased the number of results predicted within .5 goals and 1 goal by about 7 percent for the multivariate linear regression model and by about 8 percent for the neural network model. This is most likely due to the fact that a large portion of soccer games end with a goal difference within 1 and -1, so pushing predictions farther from the mean reduces the number of predictions close to this range. The highest final predictive accuracy obtained by any of the models, therefore, was 45.5 percent. This compares favorably to the bivariate poisson distribution model discussed earlier in the paper, which claims "The odds of the model outperformedbookmakersfi odds 42.22 percent to 39.23 percent in predictive accuracy" [4]. Therefore, it appears our model has predictive accuracy significantly greater than some odds or models.

## 4.3 Limitations

In retrospect, it seems that the models' biggest flaws are both their lack of ability to incorporate the z3 and z4 features into their prediction. Going into the project, I had hypothesized that this should be the biggest indicator of which team would win, regardless of their play during their last five matches. This hypothesis was supported by the fuzzy neural network model we attempted to imitate as well. Furthermore, the repeated success of a handful of teams over the rest of the league would seem to indicate that a team that beats another team multiple times would continue to do so. This is due in large part to the fact that certain teams have far more money than others to spend on players, facilities, etc., and this does not typically change from year to year. Neither model in this project seemed to indicate that previous matches between teams had a significant affect on the match prediction. Common sense would therefore seem to indicate significant issues with the implementation of the model in our project. In future implementations, we could try a different way to handle cases where there are not two previous meetings between the teams. We could, for example, give the benefit of the doubt to the team that has been in the league the longest, as we did with the last5 function. It may also be that the hole in our data in the early two thousands could be significantly affecting that feature. The csv files for those years were significantly different and could not be processed with the same function used to load the data from the other years. This is another problem with our model, as it is based on data that is incomplete and skips three years. The model we attempted to imitate had a much smaller sample size (about 1000 compared to about 8000 for ours), but presumably had accurate data points for each sample. Finally, more work with tuning neural networks would be extremely helpful, as this is a first attempt at any sort of data analysis to this degree. More experience with setting up models and more knowledge of advanced statistics would presumably greatly strengthen a model such as this. It may also be helpful to take each season independently, as teams undergo significant change over the course of the offseason. In the model in this project, a team's first game of the season is predicted using the last five games from last season, which could have been a very different set of players. In the future, it could be useful to run comparisons between our project model and a model that ignored the first five games of the season for each team so as to "get a feel" for the way a team starts out the year instead of assuming they will pick up right where they left off from the season before. Further improvements could take into account other factors such as injuries, suspensions, refereeing or weather. Clearly, we were unable to replicate all facets of the other model. The genetic tuning was extremely complicated and tough to figure out, and the neural network had more sophisticated layers and tuning. Our sample size was greater, but the data was arguably less precise due to not having previous results for every game. Furthermore, breaking results down into categories using fuzzy logic could potentially enable the model to make more accurate predictions. The fuzzy logic is less specific than a prediction across a continuous range, but this could be a good thing when tuning or fitting a model because each permutation of potential inputs would be more likely to have been seen before a prediction is made. Another potential improvement could be to simply include more features. As indicated in the

introduction, the financial value of a team is a significant indicator of a team's success, particularly in the premier league. According to [7], Chelsea is worth 631 million Euros, while Huddersfield Town is worth just 58 million. This great divide in value is seen consistently across top European leagues and could be a significant indicator of the results of matches due to the importance of being able to buy the best players. While this statistic could be partially accounted for in the previous two meetings between teams, it could be valuable to include it as its own feature in the predictive model. A potential problem would be the skyrocketing value of clubs over the last twenty years, as it would be difficult to scale the data appropriately before fitting the model. Another feature to consider would be to augment the last five played and previous two meetings features in this model by taking into account the strength of schedule. For example, beating a team that was on a win streak would be more valuable than beating a team that had lost its last five games. This could be incorporated as a multiplier in the last5 and prevMeetings functions. In this way, more information could be included in the predictive model.

## 5 CONCLUSION

In the future it would be interesting to compare an improved model to gambling lines to see how different the results are and whether there were any patterns to predicted results differing from betting lines and actual results. It could be possible, for example, that if there is a big enough discrepancy between a model's predictions and a gambling line that it would actually be worth betting. As it is, the model is not nearly accurate enough to be used to reliably predict results of sporting events, as it only out predicts random chance by twelve percent. An improved model could potentially be used to set betting lines or to inform how to beat betting lines. As this was a first attempt at this sort of data analysis and the problem is quite complicated, this was a reasonable result. However, this is again not a model to be used for practical applications. It would be interesting to see how much more of the model in the example from paper 2 could be recreated. If a more accurate predictive model like this was created, it could potentially be used by soccer teams in order to predict which matches are more likely to be won or lost so as to determine which matches would be best to rest key players. A match that is predicted to be a three goal win for your team, for example, would be a much better game to rest a key player than a game predicted to be a draw. The tactics of a team could change based on the prediction, such as a coach playing more defensively in a game their team was predicted to lose, and instead bank on tying in order to get some sort of positive result out of it. Furthermore, the model could be applied to other sports quite easily. By properly scaling data, data from sports such as football and basketball could put into the model and used in the same way. If the range for the goal difference of each game was changed, the theory behind the model could be tested for other sports as well. A future project could be to examine the difference in the correlation of past results in basketball or football versus soccer. Another potential change to this model would be to predict the number of goals each team will score instead of only predicting the goal difference. The model could be quite similar. Instead of the last5 function returning only the goal difference, it could return a list of lists, each

of which stores the number of goals scored and the number of goals conceded for each of the last five matches. Next, the prevMeetings function would change to the same format. The sampler function could return a prediction for the number of goals scored and the number of goals conceded by each team using a regression model or neural network. Next, these predictions would be fed into another regression model or neural network that predicted the number of goals scored or conceded by either team. These second layer models would combine a team's goals scored prediction and the opposing team's goals conceded prediction in order to predict the number of goals scored. Similarly, they would combine a team's goals conceded prediction and the opposing team's goals scored prediction in order to predict the number of goals scored. Both a multivariate linear regression model and a neural network could be used for this format. These predictions could give a more accurate representation of a game and incorporate some of the variables that a "per possession" model would use, while still using the same data as before, albeit in a slightly different way. By diversifying what the model is able to incorporate but still using simple and accurate data, the new model could obtain the best qualities of the previous models described without adding significant computational time. Each function might be expected to take longer, but these changes would not alter the big-O worst case time complexity of the model in general. The advantages of this sort of model's prediction is that it could be used to make more specific types of gambling lines, such as over/unders. Furthermore, it could better inform a coach on tactical decisions about how aggressive or defensive to align his team. The models described in this paper could make a great starting point for predictive modeling of any kind, as the models included are extremely flexible. Although the majority of the work ended up being actually finding, parsing, and formatting the data correctly, the methods used to analyze the data would work with any big sets of data in a dataframe format.

## 6 ACKNOWLEDGEMENTS

The author would like to thank Juliette Zerick for their help throughout this course.

## REFERENCES

- [1] H.P.H. Eggels. 2016. Expected Goals in Soccer: Explaining Match Results using Predictive Analytics. (2016). Retrieved Oct 30, 2017 from <https://pure.tue.nl/ws/files/46945853/855660-1.pdf>
- [2] Football-Data. 2017. Football-Data. (2017). Retrieved Dec 1, 2017 from <http://www.football-data.co.uk>
- [3] Jurgen Gerhards. 2016. Who wins the championship? Market value and team composition as predictors of success in the top European football leagues. (2016). Retrieved Dec 5, 2017 from <http://www.tandfonline.com/doi/abs/10.1080/14616696.2016.1268704>
- [4] A. Groll, T. Kneib, A. Mayr, and G. Schauberger. 2016. On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016. (2016). Retrieved Nov 1, 2017 from <http://eprints.kingston.ac.uk/39162/1/MathSport2017Proceedings.pdf#page=166>
- [5] Josh Lipe-Melton. 2017. Big Data Applications in Team Sports Predictive Analytics. (2017). Retrieved Dec 5, 2017 from <https://github.com/bigdata-i523/hid105/blob/master/paper2/report.tex>
- [6] A. P. Rotshtein, M. Posner, and A. B. Rakityanskaya. 2005. FOOTBALL PREDICTIONS BASED ON A FUZZY MODEL WITH GENETIC AND NEURAL TUNING. (2005). Retrieved Oct 30, 2017 from <https://link.springer.com.proxyiub.uits.iu.edu/content/pdf/10.1007%2Fs10559-005-0098-4.pdf>
- [7] TransferMarkt. 2017. TOTAL MARKET VALUE TREND OF ALL CLUBS OF PREMIER LEAGUE. (2017). Retrieved Dec 5, 2017 from <https://www.transfermarkt.com/premier-league/marktwerteverein/wettbewerb/GB1>

- [8] UFLDL. 2017. Optimization: Stochastic Gradient Descent. (2017). Retrieved Dec 9, 2017 from <http://ufldl.stanford.edu/tutorial/supervised/OptimizationStochasticGradientDescent/>
- [9] Wikipedia. 2017. Rating Percentage Index. (2017). Retrieved Oct 30, 2017 from [https://en.wikipedia.org/wiki/Rating\\_percentage\\_index](https://en.wikipedia.org/wiki/Rating_percentage_index)
- [10] Albrecht Zimmermann and Jesse Davis. 2013. Machine Learning and Data Mining for Sports Analytics. (2013). Retrieved Oct 30, 2017 from <https://lirias.kuleuven.be/bitstream/123456789/424505/1/CW650.pdf>

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
=====
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
report.bib:76: @misc{ExpectedGoals,
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-16 09.31.49] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflat
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 0.9s.
```

```
=====
Compliance Report
```

```
=====
name: Lipe-Melton, Josh
hid: 105
paper1: 100% October 27, 2017
paper2: 100% November 6, 2017
project: Dec 05 17 100%
```

```
=====
yamlcheck
```

wordcount

---

7  
wc 105 project 7 6686 report.tex  
wc 105 project 7 6629 report.pdf  
wc 105 project 7 300 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

passed: False

find input{format/final}

---

4: \input{format/final}

passed: True

floats

---

figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0

True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
True : check if all figures are referred to: (refs >= labels)

Label/ref check  
passed: True

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

```
-----
```

```
passed: True  
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

```
-----
```

```
passed: True
```

# Big Data Analytics in Indian Premier League

Swargam, Prashanth  
Indiana University Bloomington  
107 S Indiana Ave  
Bloomington, Indiana 47408  
pswargam@iu.edu

## ABSTRACT

Cricket is one of the most admired sports across the globe. Indian Premier League is one of the professional cricket leagues conducted by Board of Cricket Control India in the months of April and May. This league is famous for its diversity of players and breath-taking cricket match endings. The factors of winning change for each moment as the game progresses. As there are many players and franchises involved in the game, these factors for winning changes for each team. Data related to each player is required to analyse his performance and predict his future scope in team. Data related to factors of winning is crucial and can be analysed for predicting the results of the game. This analysis would help the team management, league administration to wisely chose the players and modify rules according to the impact of each decision. Data related some of the important factors which plays major role in deciding the match winner are analysed. Their impact is predicted and compared with the actual results. Impact of these factors are studied for each individual team and individual season of this cricket tournament. Impact of each factor is plotted and its impact in next season is predicted.

## KEYWORDS

Big Data, Cricket, Indian Premier League, i523

## 1 INTRODUCTION

Fast paced games are gaining more importance in near future. This because there are many factors which contribute to the result of the game. These factors are minor but could change the results of the game dramatically. Indian Premier League is one such type of cricket league where there are a lot factors which have their influence on the results of the game. These factors are though minor or major, will have bigger part in deciding the results of the game. These factors from the previous games can be utilised wisely to predict their influence in the upcoming matches. These factors can be quantitatively represented in the form numbers, graphs or Booleans. This quantitative representation of data related the factors can be analysed using various analytic techniques to predict their impact on the game.

However, Analytics is a good way to go about this prediction, but there are several problems which should be addressed. Considering the role of batsman, it will be having parameters like balls faced, dot balls, number of boundaries, strike rate etc. Considering the role of bowler, there are various parameters like matches played, overs bowled, economy rate. There are many similar kinds of roles in the game and above-mentioned parameters are specific to one player playing only one role in the team. According to, around 500 players play for each season of cricket. These 500 players will be

filtered on various factor from the pool of nearly 5,31,253 cricket players across the globe. These players can play any of the role or play multiple to roles to contribute to the result of the match. These players and cricket matches produces large amount of data which when analysed to produce structured data and analytics. Hence, there is good scope of analytics ad big data in this sport.

The data produced by the matches happening in Indian Premier League can be used to fit in mathematical models. These mathematical models are then used to study the nature and trends of the factors which influence the results of the game. Extending this model to the known values of the input factors can produce the predicted values of the impacts of these factors. Models like Linear regression, polynomial regression, radial-basis approach can be applied to do these kinds of predictive analysis.

## 2 PROBLEM STATEMENT

There are various factors influencing the results of the game. As part of this analytics, data related to four of the most influencing factors is gathered and modelled for analysis. This data was available in raw formats which requires some amount of modelling for predictive analysis. The modelled data is used for building a mathematical model which would fit closely to the trend of these factors in matches played in all the past seasons. A part of data is assumed to be unknown. This unknown part of data is predicted by using the fitted mathematical models. Results obtained by these predictions are compared with the actual results from the data source. Impact of these factors are calculated to the ratio of one. These data is analysed for each independent team and each independent season.

The report is in regard to the predictive analysis conducted on only five of the most influencing factors in the game. There are other factors in the game which might influence the result of the game. This predictive analysis will only be considered reliable only if the predicted values of the results will have high accuracy with respect to the actual results of the available data. The data is divided into two parts. All the available data is sorted with respect to date. The latest match comes later in the dataset and the earliest first. The first part of data is used to train the mathematical model. The parameters in the later part of the dataset are used to predict the result of the match. These predicted results are then compared with the actual results in the datasets to determine the accuracy of the predictive model which is used to build the analytics. This analysis produce valuable insights on the influence of these factors and the mathematical model.

## 3 SCOPE

The scope of the analysis is:

1)The analytics uses the data for only five factors. These five factors are namely toss, Batting position, Range of score, portion of runs in boundaries.

2)The data is collected for all the seasons completed for this tournament. However, this data is sorted with respect to date and partitioned into training and testing data sets for calculating the accuracy of the model.

3) The values of these factors are represented in usable data formats like range, Boolean, integers for analysis.

## 4 FACTORS IN CONSIDERATION

### 4.1 Batting Sequence

Order of batting is considered as one of the factors in consideration. Batting order is one crucial parameter which depends on various other factors of the game. Some of these parameters are the status of the pitch for the game, climate conditions of the game, previous statistics of the game and the history of the team in similar situations. The toss winner will have the privilege to decide the order of the batting. As this factor is conglomeration of various other factors stated above, batting order is considered for the analytics. This data can be represented in the form of Boolean. Where true Boolean indicates that the team referring to the statistics have batted first in the game. False indicates that the team referring to the statistics have batted second in the game. This Boolean value depends on the values in for toss winner and toss decision.

### 4.2 Total Score

Score indicates the total number of runs scored by the team in any match. Score of the team depends on various other parameters of the game like team statistics and composition, impact of the opponents, and situation of the match. This parameter can be calculated from other values of extra runs, scored runs. This categorized into four categories. This first category of the innings scored not more than 100 runs. The second category of the innings scored more than 100 and less than 150 runs. The third category of the innings scored more than 150 and less than 200 hundred runs. All the other innings which scored more than 200 are categorised into fourth category. This categorization is done in accordance to the range of scores. The least scored innings were given least category value. The highest scored innings were given highest category value.

### 4.3 Score Composition

Composition of scored runs. The runs are majorly scored in the form of boundaries, players individual running, and the extra runs given by then bowling team. As IPL is a T20 game which is played for short duration of time, scoring runs quickly at right time is crucial factor. Boundaries contribute to runs scored in the form of fours and sixes in the game. This is the easiest way to quickly score the runs. Team scoring high majority of runs in the form of boundaries have higher chance of imposing a higher target to the opponents or chasing down the target imposed by the opponents. Hence, this parameter is considered for analysis. This value for this parameter is a Boolean. This value is set to true, if most of the runs scored by any team in any innings are from boundaries and vice versa.

### 4.4 Toss

The batting sequence is decided by the winner of the toss. Winner of the toss will have the initial upper hand in the game to decide the sequence of the game. The winner's decision will vary on various other factors of the game like, duration of the match, pitch behaviour throughout the game, statistics of the game. These various factors play an important role in deciding the toss winner's decision. The value for this parameter is Boolean. The true value of the parameter indicates the team has won the toss and the false value indicates the team has not won the toss.

## 5 DATA MANIPULATION

### 5.1 Team data

There are thirteen teams participated in IPL which was held for nine seasons. Each team was having its team name and teamId. These details are taken from the data source team.csv. Python's csv module is used to read the data from this csv file. As this data represents a key value pair, csv module's dictreader method is used to read the row in these files. Using this method, a dictionary was created which consisted of the teamId as keys and team name as value objects.

### 5.2 Match data

Details pertaining to a specific match are published to the Match.csv file. These details include host team name, guest team name, toss winner id, match winner id, decision of the toss, win type. This data was used and modified for calculating the impact of each factors stated in the factors description. Python's pandas dataframes were used to read the data from these csv files. All the missing values in the dataframes are replaced with 0 to ease the complexities that arise with null values. For each value in the team dictionary, the teamId is matched with the opponent team id column and team name id column in the match.csv. Python's operator module is used to obtain the or condition between these values. Dataframe is modified with the given conditions. This dataframe is converted to list. This way, list of matches played by a team is defined and stored into a list.

From the dataframe which contains the list of matches played by the team, column matchwinnerid is used to define if the match winner. A new dataframe is created with a condition if the matchwinnerid's value in the column is equals to the id of the current team. If the above-mentioned condition is true, then this value is added to the dataframe, else the value is removed from the dataframe. This way, list of matches won by a team is determined.

A new dataframe is created to store the Booleans of the toss decision. From the teamdata dataframe, the column toss winner id is used to determine the toss winner for each match. If the id value in this column is same as the team id of the current team, Boolean true is appended to the toss list. Else, Boolean false is appended to the toss list. This list contains only Booleans.

### 5.3 Ball-by-ball data

Ball by ball analyses will be required to calculate the scores for each ball. These details will include the number of runs scored in the ball, extra runs scored, bowler details, batsman details, over details. This

data is extracted from the ballbyball.csv file. This file contains ball by ball analysis of all the matches. This data is sorted according to the match id and is extracted for further analysis. Pandas readcsv method is used to read the csv file. A new data dataframe called balldata is created.

Balldata csv is used to for calculating the runs scored by a team in a specific match. The balldata dataframe is filtered for useless columns and null values. All the null values are replaced with 0 to ease the complexities which comes with usage of null values. Balldatadfis team batting id and match id are used to calculate the runs scored by a team in a specific match. This data frame is modified such that, the values in the match id column is equal to the current match id and the values in the team batting id is equal to the current team id. Python operator module is used to achieve the and condition in the above case. The modified ball data data frame is used to calculate the score. The sum method on the dataframe is used from the pandas library. This score is categorised into four categories. The first category included the score which are less than hundred. The second category included the scores which are between hundred and one hundred fifty. The third category included the scores which are between one hundred fifty and two hundred. The fourth category of scores contain scores which are above two hundred.

A new list is created which stored the category of the scores. If the score falls in first category, then an integer 1 is appended to the list. If the score falls in the second category, integer 2 is appended to the list. If the score falls in the category falls in the third category, an integer 3 is appended. If the score falls in the category four, an integer 4 is appended to the list. The other factor which was stated in the factors in consideration teamfis score composition. It is highly probable that a team scores a high total or chase down the target imposed by the opposition team quickly is the majority of runs are scored in the form of boundaries. The ball data dataframe which is created in the previous case is utilised with other conditions to calculate the contribution of boundaries to the total score. This dataframe is filtered with the match id and team id condition to obtain the current match and current batting team. This is again filtered for all the scores that are having values either four or six. The minified frame is filtered for all values of four and summed. The same is repeated with the sum of six values. Adding together these two sums will give the total amount of runs scored in form of boundaries.

A new list is created for storing the Booleans related to the contribution of boundaries to the score. This list is appended with value true, if the contribution of boundaries is more than other forms of runs, false if, the contribution of boundaries is less than the contribution of other forms of runs.

The lists which are created for each factor are used to define factorsfi impact on the gamefis result. These lists contain the values in the form of integers and Booleans which are obtained by using the existing parameters. These lists are created for each value in the team.csv file. Thus, they are independent to each team. A new DataFrame is created with these lists as columns in the frame. For each team, these frames are stored into another csv file using the pandas write csv function with the file name same as the value in the team dictionary.

## 6 PREDICTIVE ANALYSIS

The factors and their values are written in to a csv file which contains the factor names as the columns headings and their respective values in the rows. Each team has its independent statistics files mentioned in the above statement. These files will be used for conducting the predictive analysis. For each value in the team dictionary, these csv files which are specific to the team are read using the pandas csv reader function. There are two types of columns in these statistics file. The first type is predictors and the other type is targets. Columns related to the factors which impact the results are considered as the predictor columns, Columns related to the result of the match are considered as the targets column. Columns battingfirst, boundaries majority, wontoss, scorerange are predictors. Wonmatch column is target in the given scenario.

The data available in the statistics files is split into test and train sets. This is done by the function train test split in the sklearn library in the python. The data is split in the ratio of 3 is to 2. Sixty percent of the data is used as training set. Forty percent of the data is used as testing set. The predictors and target of the training set are used to build the mathematical model. The predictors of the testing set are used to predict the values of the targets. These predicted values are compared with the actual target values. This comparision is used to determine the accuracy of the prediction.

### 6.1 Implementation

Decision trees and Random forest are used in making the predictions. Decision trees are tree like structures which are built based on the values of the parameters. These trees are useful in defining the probability of the target value being attained. Trees are build with various stems which are drawn from conditional statements. Decision trees has three kinds of elements. The first element i.e., are the decision elements which refer to the block which checks for the condition or logic of the tree. Chance elements are the elements which occur depending on the condition or logic of the function. End elements are the results or outcome of the decision tree. These are basically leaf nodes of the tree. These nodes represent the result. Random Forests are conglomeration of decisions from various decision trees. In random forest approach, a dataset is divided into various subsets which will have some or all the input parameters as the decision makers and some or all the data which are the values for the parameters. Each subset is used to build a tree by using principles of decision tree. The predictions from these prediction trees are used in determining the final value. When a given set of input parameters are giving, they are predicted with all the decision trees developed by the forest. The outcome from all the decision trees are noted. The majority of these outputs is decided as result. This way, the errors which might arise in using only one decision tree can be eradicated. An error from one model of decision tree will be dominated by the results from all the other decision tree.

Random forest classifier from the module sklearn is used to build the various decision trees and predict these values. Classifier type object is instantiated in the code/cite. This object is assigned with the RandomForestclassifier and an attribute called n estimator. The variable n estimator will define the number of decision trees to be build for the analysis. Fit function from the sklearn module is used to develop the model for the random forest algorithm. Fit function

will take the training parameters and training targets as inputs. These are divided into fifty subsets in this scenario. Predict function is used to predict the value of the target given test predictor variable as inputs.

## 6.2 Accuracy

Generation of only one decision tree as the model for the given training data would produce erroneous results. Using the random forests, fifty decision trees are built to predict the correct value of the target. This way, errors produced by one of the decision tree will be corrected by the predictions from the other trees.

In the graph 3, the accuracy of using various number of decision trees are plotted against the number of decision trees. It can be observed that using less than five decision trees, the accuracy for the prediction for all the teams is around sixty percentage. As the number of decision trees increased, the accuracy of all the prediction is increased by at least ten percent. The hundred percent accuracy is because, those teams have played less number of games.

## 7 IMPACT OF FACTORS

This indicates the contribution of each factor for predicting the result of the game. These values will determine the probability of the value of result on any given values for the input variables. In the first step, the distinct values of the target value are noted. For all the distinct values of an input parameter from the set of values of one of the input parameter, the probability of different kind of results are calculated. This calculation is repeated for the distinct values in the set of input parameter and summed at the end. This produces the importance of the feature. The above procedure is repeated for all the decision tree and all the value are averaged for getting the overall value of the importance feature for that feature. This procedure is repeated for all the input parameters. This will give the contribution of each parameter to the result values.

### 7.1 Batting First

The first importance feature for all the teams are plotted in the graph 4 . It is clear from the graph that the importance feature batting first is highest for Rising Pune super giants. This indicates that this team have won most of their previous matches with while batting first in the game. While the Chennai Super kings and Kochi Tuskers are also high, but they are less than half. This indicates that batting first in the match will be a favourable condition for the above mentioned teams. For all the other teams except Kings xi Punjab, Pune warriors and Sunrisers Hyderabad, the batting first importance factor is nearly around 0.1. This implies, out of all that matches which were present in the training set, these team batted first for only ten percent of the games.

However, the total number matches should also be considered while validating this condition. Though the value for Kochi Tuskers is high, this might also be because they have played less number of matches and they have batted first in all the winning matches.

### 7.2 Score Composition

Score Composition composition is the combination of different ways of getting runs. For any high scoring and successful game, a team will have to score runs at faster rate. Hence, scoring runs in

form of boundaries will help a team a lot in turning the match in their favour. In the graph 5 , Contribution of this factor against each team is plotted. This graph summarises the contribution towards of the score composition factor towards the factor. This factor is very high for Kochi Tuskers team, because they have played considerably less number of games.

A factor around 0.2 to 0.3 seems around the average value for all the teams. It is clearly seen from the graph that this factor is high for Rising Pune Supergiants. That means, out of all the matches this team have won in all the season, in most of the matches, this team have scored most of their runs in form of boundaries. Then, Rajasthan royals and Delhi Daredevils are having next higher values. A lower value of this factor means, that out of all the matches which this specific team have won, they have scored most of them in form of another form. Sunrisers Hyderabad and Royal challengers are one such team.

This factor is calculated independent of the team composition. This factor can be normalised with respect to the team composition.

### 7.3 Score Range

Score is the total number of runs scored by a team in a specific innings of the match. This factor is categorised into four categories. Each category has a range of value for runs. Based, on this runs, the team is classified into categories. The first category of team will have scored less than hundred runs. The second category of team have scored less than one hundred fifty. The third category of team have scored less than two hundred runs. The remaining teams comes under the fourth category.

This categorisation is used as one factor in determining the results. From the graph, it can be seen that this value is very high for some of the teams and considerably less for other teams. A higher value of this factor means that out of all the matches the specific team has won, most of the matches they have scored the runs in the higher categories ,or out of all the matches they lost, they have scores in the lower categories of the score.

From the graph 7 ,It is clear that, teams Kolkatta Knight Risers and Sun Risers Hyderabad have a value around 0.6. This implies that the wonmatch of column for this teams was mostly decided by the score range category column. For the teams like Gujarat Lions and Rising Pune Super giants, this value is low. It can be inferred that the wonmatch column for this table is mostly decided by other columns in the team statistics table.

### 7.4 Toss

Toss is another important factor considered for this analysis. The winner of the toss has the power to decide the sequence of the match. This decision of the winning captain will effect the results of the match. Given this opportunity, any captian would take decision in favour their side. Hence, toss is one important factor in deciding the results of the match.

A higher value of this factor implies, that out of all the matches the specific team has won, they have also won the toss in most of the matches. It also implies that the column wontoss have contributed a large amount to the target column. A lower value of this factor implies that out of all the matches a specific team has lost, most of the matches they have lost the toss.

From the graph 6 it is clear that the this value is higher for teams Delhi Daredevils, Gujarat Lions and Royal challengers bangalore. This implies that out of the matches they have won, most of the matches they have won the toss. This implies toss is one of the important factor for this team to win a specific match. It is clear from the graph that this value is lower for teams like Kolkatta Knight risers. This implies that out of all the matches won by this team, they have won toss in less number of matches. This implies that toss is not one of the important factors for this team to win.

For the other teams, this value is around 0.15 which is considered as average contribution. For these teams, toss decision have a fair impact on the decision of the match. The wontoss column in the team statistics have contributed fair amount to the target column.

## 8 STATISTICAL ANALYSIS

This analysis will give the statistics of each factors and their importance in the previous matches for each team. This is done by gathering the data from the team statistics which is prepared as part of the first step in the predictive analysis.

A dictionary of team names is prepared using the teams.csv file from the source. This file is read using the read csv method in the csv library. An empty list is created for the factors batting order, score range, toss. As there is only one kind of value for all these factors, they are grouped and studied together. The score range is studied in a different program.

These empty lists are used for storing the percentage contribution of each factor towards the results of analysis. In this case, the data is not divided into test and training sets. All the data in the data sets are taken into consideration. The values used in the analysis are the actuals value and no predictions are made in defining these values.

For every element in the team dictionary, we iterate over the specific team statistics csv file created in the first step of this predictive analysis. From these csv files , we read the columns battingsfirst, majorityscore, wontoss columns. For each columns in the csv file, we create a corresponding dataframe in the python program using the pandas dataframe. This dataframes are constructed on based on two values. The wonmatch column value and the value of the factor being studied. We append the corresponding row to the dataframe only if both the conditions are satisfied. This kind of conditional statements can be achieved by python or operator.

For every factor, now independent dataframes are defined after both the conditions are satisfied. Total matches is the length of the csv file. From the above calculate the percentage of the dataframe which we have captured with respect to the total length of the csv file. This percentage value determines the percentage contribution of the factor towards determining the winning chances of the team.

These percentages are calculated for all the teams in the team dictionary. These values are stored in the respective factors list. This list is used for plotting the bar graphs using the pythons matplotlib.

### 8.1 Analysis of individual team

From the the lists of percentage contributions from the above analysis. Analysis of each factor and their contribution towards the

individual team has been plotted in the graph. These plots are plotted against the team name and the percentage bars which show the percentage of each factor. Graph 2 has been plotted with above lists.

For the team Kolkatta knight riders, out of all the matches they have won, in forty percentage of those matches, they have score majority of their score in form of boundaries. Out of all the matches they have won, they also won toss in thirty percentage of the matches and batted first in twenty percentage of the matches. This implies that the probability of winning a match for kolkatta Knight riders team is high if they score moajority of their score in form of boundaries. Then decision of toss and sequence of matches have considerable effect in the matchesf decision.

Team Royal Challengers bangalore have followed the same trend as the kolkatta team in the analysis. Out of all the matches they have won, they also scored majority of the runs in the form of boundaries. While ,out of these matches, they won only twenty percent of the tosses, they batted first in nearly twenty five percentage of the matches. Runs have been major contributing factor in this team also.

Chennai Super Kings have majority of their contribution from the toss factor and batting sequence factor together. That implies that, from the all the matches they have won, they either won the toss or batted first in the match. This implies that the team Chennai Super kings will have to win toss or bat first to win the match.

Teams punjab and Kochi Tuskers have followed trend similar to that of Kolkatta Knight risers and royal challengers bangalore. They scored majority of their score in the form of boundaries in nearly forty percent of the matches played by them. The other factors toss and batting sequence have contributed to nearly twenty percentage of their winnings. This implies that scoring majority of runs in form of boundaries will favour these teams.

Mumbai Indians team have the second highest percentage contribution from the factors toss and batting sequence compared to other teams. This team also have third highes contribution from the score composition factor. That implies, this team have higher chance of winning given any factor. Though the value from the any one factor is against them, the other factor will over ride the effect of the previous factor.

Gujarat Lions have the highest contribution from the factor batting order. That implies, out of all the matches won by this team, they have scored majority of runs in the form of boundaries. The other two factors are considerably low. This implies that Gujarat team will have to score most of the runs in the form of boundaries to win the match.

It can be inferred from the graph that rising pune super giants have not won a match while batting first in the game. Out of all the matches won by them, they have either batted second in the game or score majority of runs in the form of boundaries.

Team Deccan chargers have almost same amount of contribution from all the three factors. This team is consistently performing in any values of the factors. They have good balance of the conditions in the previous games.

Team Pune warriors has the lowest values for the factors toss and sequence of the match. They are also have less contribution from the factor composition of score. This implies that this team is under performing in any given condition.

## 8.2 Analysis of Range of Scores

The scores were divided into categories. This analysis will contribute the percentage contribution of each range of the score to the result of the matches played by the specific team. This analysis can be used as to determine the safe score range for every team.

Team dictionary is taken from the team.csv file. The team statistics csv file which is created in the predictive analysis will be utilised for the value for range of scores. The values from the columns score range and wonmatch will be utilised. The data in this csv file is not partitioned into training and testing sets. This analysis is performed on complete data. No data is predicted in this analysis also. This is analysis on all the available data.

For every value in the team dictionary, we locate the team statistics csv file which is generated as part of the predictive analysis. Four empty lists are generated to store the number of matches won by each team with score in the given score range. Total number matches can be calculated by using the length of team statistics csv file that is being studied upon. For every value in the team dictionary, we divide the csv file into four different data frames. Each data frame corresponds to each range of scores. This partition can be achieved by using pandas data frame.

This dataframes are extracted using the two conditions. They are the value of score range must be equal to the range we are fetching data for and the other condition is the wonmatch column must be true.

After corresponding data frames are extracted from the team staistics csv, we calculate the percentage of each range data frame length against the total length of the team statistics csv. This percentage value will give us the percentage contribution of each range towards the result of the match.

## 8.3 Graphical Analysis on scores

From the graph 1, it is clear that most of the wining scores fall in the second and third category. That implies for any team to win the match, it is most likely that the team must score atleast hundred runs in the match and come under second category or score atleast one hundred fifty runs and come under third category.

Teams Gujarat lions and pune warriors have never won match scoring less than hundred runs or more than two hundred runs. This implies that the range from hundred to two hundred is the safe score range for these teams. Gujarat have won most number of matches scoring in the third category that is atleast scoring one hundred fifty runs and atmost two hundred runs. This is more safe zone for them. While, for pune warriors team they have almost similar winning percentage in both the categories. Teams rising pune supergiants and kochi have never won a game scoring more than two hundred runs. This might be bacuse, they have not scored more than two hundred runs in any given match or they have not won in the matches in which they scored more than two hundred runs. That implies that they are having a safe scoring in the range of one hundred fifty and two hundred.

Sunrisers Hyderabad team has more contribution from the the second and third categories of scoring in the game compared to the other two teams. This indicates that they are consistent in this category of scoring than other teams. They do not have major difference in contribution from these categories. This indicates that

they are consistently scoring around one hundred fifty mark score which makes the assumption that this team have good batting side. From the statistics on this team it is clear that they have also won matches scoring less than hundred. It is clear that they are also having a good bowling side as well.

Team Royal Challengers bangalore has highest contribution from the fourth category of scoring when compared to other teams. That implies, the team royal challengers banglore have highest probability of winning the matches , if they are scoring more than two hundred runs in the match. Then their next contribution comes from the score range of third category. That implies this team has a good batting side. Because they are having high scoring percentages from the third and fourth categories.

Following the royal challenger banglore team is Chennai super kings team. This team have contributions from all the four categories of the scoring range. This implies this team is fairly consistent in both the categories of the game.

Team Mumbai Indians have the highest contribution from the third category of the scoring range after Gujarat lions. They are also having contributions from all the categories of range of scores. This will clearly show that Mumbai Indians team has an inconsistent batting team. If these totals are from the second innings, it can also be assumed that whenever their batsman failed to score many runs, bowler won the match.

Kolkatta team has the highest contribution from the first category of scoring after Kochi team and Pune super giants. This implies that the above three teams have a good batting side. Scoring less than hundred and winning match is sight of team having a good bowling side. Upon having good contributions from category one, kolkatta knight riders also has good contributions from the other three categories as well. This implies, that this team not only has good bowling side,it also ahs a good batting side as well.

## 9 CONCLUSION

Predictive Data analytics have provided promising solutions to various problems in wide variety of fields. As part of this study, predictive and statistical data analytics on data related to a cricket League, Indian Premier league is conducted. As part of predictive analytics, the available data is split into training and testing data sets. The values from the model are used to predict the target value. These values are compared to the original values for accuracy. Method of improving the accuracy of model is studied. This study would be useful to determine the impact of a factor on the result of the game. This analysis would help in predicting the results of the matches acuurately with the model developed from the training data.

Statistical analysis on the same data is conducted to get the details related to the impact of a factor quantitatively on a specific team. This analysis pertains to the available data and no predictions are done. This kind of analysis would be useful in determining the strengths of individual team. This kind of analysis can be conducted to the nature and strengths of each team.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this report.

The author would like to thank assistant instructors for their support in completing this project.

The author would like to acknowledge that the base data for the analysis is provided by [1].

## REFERENCES

- [1] HarshaVardhan. 2016. Indian Premier League. (2016). <https://www.kaggle.com/harsha547/indian-premier-league-csv-dataset/data>

[Figure 1 about here.]

#### LIST OF FIGURES

1	Score Staistics	10
2	Other Factors Staistics	10
3	Tree number vs Accuracy	10
4	BattingFirst	10
5	Majority of runs in boundaries	10
6	Toss	10
7	Range of Scores	10



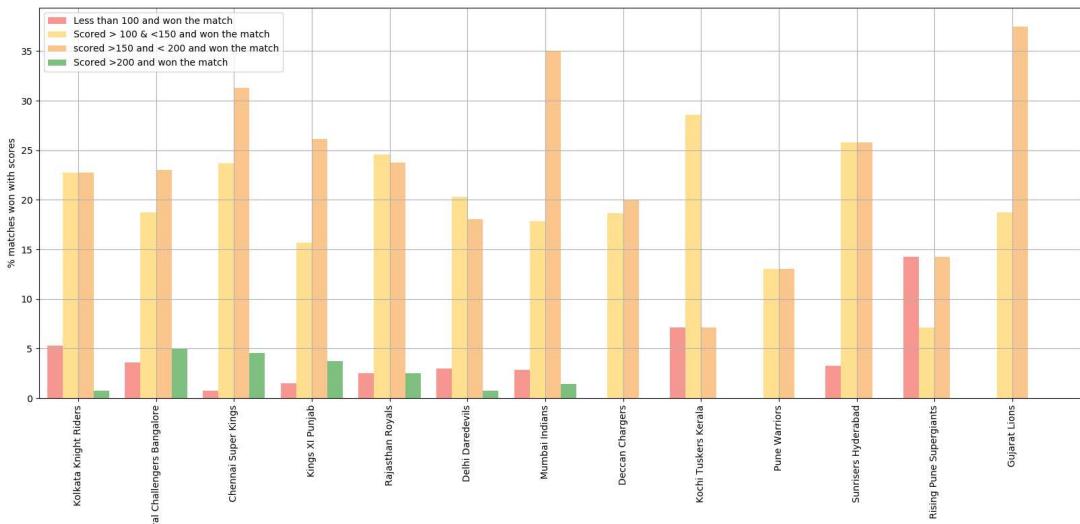


Figure 1: Score Staistics

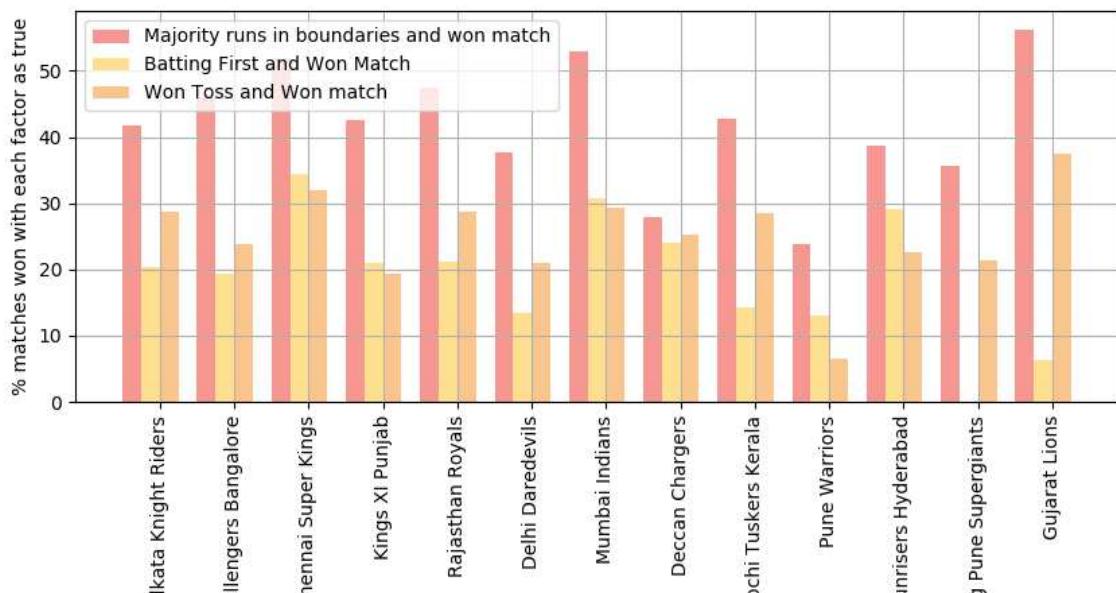
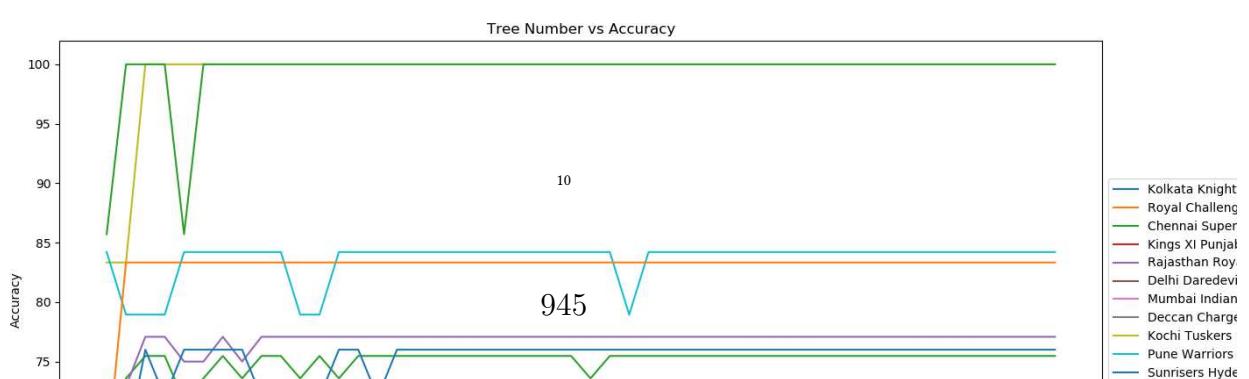


Figure 2: Other Factors Staistics



## bibtex report

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

## bibtext \_ label error

bibtext space label error

## bibtext comma label error

# latex report

[2017-12-16 09.34.10] pdflatex report.tex

```

Missing character: ""

bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Float too large for page by 1818.09471pt.
Typesetting of "report.tex" completed in 1.6s.
./README.yml
 33:75   error   trailing spaces (trailing-spaces)
 36:81   error   line too long (82 > 80 characters) (line-length)
 36:82   error   trailing spaces (trailing-spaces)
 40:81   error   line too long (89 > 80 characters) (line-length)
 53:81   error   line too long (92 > 80 characters) (line-length)
 54:81   error   line too long (95 > 80 characters) (line-length)
 54:95   error   trailing spaces (trailing-spaces)
 55:81   error   line too long (91 > 80 characters) (line-length)
 56:81   error   line too long (91 > 80 characters) (line-length)
 56:91   error   trailing spaces (trailing-spaces)
 57:81   error   line too long (94 > 80 characters) (line-length)
 58:81   error   line too long (96 > 80 characters) (line-length)
 59:81   error   line too long (99 > 80 characters) (line-length)
 60:81   error   line too long (100 > 80 characters) (line-length)
 61:81   error   line too long (100 > 80 characters) (line-length)
 62:81   error   line too long (99 > 80 characters) (line-length)
 62:99   error   trailing spaces (trailing-spaces)
 63:81   error   line too long (99 > 80 characters) (line-length)
 64:81   error   line too long (101 > 80 characters) (line-length)
 65:81   error   line too long (103 > 80 characters) (line-length)

```

---

### Compliance Report

---

```

name: Swargam, Prashanth
hid: 228
paper1: Oct 20 17 100%
paper2: Nov 06 17 100%
project: Dec 04 17 100%

```

yamlcheck

---

wordcount

---

10  
wc 228 project 10 6689 report.tex  
wc 228 project 10 6475 report.pdf  
wc 228 project 10 24 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

6: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

136: In the graph \ref{f:treenumvsaccuracy}, the accuracy of using various number of decision trees are plotted against the number of decision trees. It can be observed that using less than five decision trees, the accuracy for the prediction for all the teams is around sixty percentage. As the number of decision trees increased, the accuracy of all the prediction is increased by at least ten percent. The hundred percent accuracy is because, those teams have played less number of games.

144: The first importance feature for all the teams are plotted in the graph \ref{f:BattingFirst} . It is clear from the graph that the importance feature batting first is highest for Rising Pune super

giants. This indicates that this team have won most of their previous matches with while batting first in the game. While the Chennai Super kings and Kochi Tuskers are also high, but they are less than half. This indicates that batting first in the match will be a favourable condition for the above mentioned teams. For all the other teams except Kings xi Punjab, Pune warriors and Sunrisers Hyderabad, the batting first importance factor is nearly around 0.1. This implies, out of all that matches which were present in the training set, these team batted first for only ten percent of the games.

- 150: Score Composition composition is the combination of different ways of getting runs. For any high scoring and successful game, a team will have to score runs at faster rate. Hence, scoring runs in form of boundaries will help a team a lot in turning the match in their favour. In the graph \ref{f:BoundariesMajority} , Contribution of this factor against each team is plotted. This graph summarises the contribution towards of the score composition factor towards the factor. This factor is very high for Kochi Tuskers team, because they have played considerably less number of games.
- 164: From the graph \ref{f:scorerange} ,It is clear that, teams Kolkatta Knight Risers and Sun Risers Hyderabad have a value around 0.6. This implies that the wonmatch of column for this teams was mostly decided by the score range category column. For the teams like Gujarat Lions and Rising Pune Super giants, this value is low. It can be inferred that the wonmatch column for this table is mostly decided by other columns in the team statistics table.
- 172: From the graph \ref{f:Wontoss} it is clear that the this value is higher for teams Delhi Daredevils, Gujarat Lions and Royal challengers bangalore. This implies that out of the matches they have won, most of the matches they have won the toss. This implies toss is one of the important factor for this team to win a specific match. It is clear from the graph that this value is lower for teams like Kolkatta Knight risers. This implies that out of all the matches won by this team, they have won toss in less number of matches. This implies that toss is not one of the important factors for this team to win.
- 191: From the the lists of percentage contributions from the above anaylsis. Analysis of each factor and their contribution towards the individual team has been plotted in the graph. These plots are plotted against the team name and the percentage bars which show the percentage of each factor.Graph \ref{f:otherstats} has been plotted with above lists.
- 231: From the graph \ref{f:scorestats}, it is clear that most of the wining scores fall in the second and third category. That implies

for any team to win the match, it is most likely that the team must score atleast hundred runs in the match and come under second category or score atleast one hundred fifty runs and come under third category.

```

272: \begin{figure} [!ht]
273: \centering\includegraphics[width=\columnwidth]{images/scorestatics.png}
274: \caption{Score Staistics}\label{f:scorestats}
276: \centering\includegraphics[width=\columnwidth]{images/otherstatics.png}
277: \caption{Other Factors Staistics}\label{f:otherstats}
279: \centering\includegraphics[width=\columnwidth]{images/treenumvsaccuracy.png}
280: \caption{Tree number vs Accuracy}\label{f:treenumvsaccuracy}
282: \centering\includegraphics[width=\columnwidth]{images/BattingFirst.png}
283: \caption{BattingFirst}\label{f:BattingFirst}
285: \centering\includegraphics[width=\columnwidth]{images/BoundariesMajority.png}
286: \caption{Majority of runs in boundaries}\label{f:BoundariesMajority}
288: \centering\includegraphics[width=\columnwidth]{images/Wontoss.png}
289: \caption{Toss}\label{f:Wontoss}
291: \centering\includegraphics[width=\columnwidth]{images/scorerange.png}
292: \caption{Range of Scores}\label{f:scorerange}

```

```

figures 1
tables 0
includegraphics 7
labels 7
refs 7
floats 1

```

```

False : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
False : include graphics passed: (figures >= includegraphics)
True : check if all figures are refered to: (refs >= labels)

```

```

Label/ref check
passed: True

```

```

When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction

```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

```
non ascii found 8217
```

```
non ascii found 8217
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
=====
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
=====
```

```
passed: True
```

# TBI: A Data Driven Journey Beyond Contact Sports that Puts Data In The Driver's Seat

Jeffry L. Garner

Indiana University

Online Student

jeffgarn@iu.edu

## ABSTRACT

The data journey into concussions starts with the confluence of contact sports, long-term neurological diseases, and well known athletes. Lots of fascinating technologies to help the hockey player, football player and others that play contact sports. And hey, sports matter! But the data journey leads down other roads. The Data Scientist has an opportunity to help the athlete but also an opportunity to help many others as well. This data journey is not well known and has far less panache but is important nonetheless. This road leads to military veterans, those injured in auto-accidents, and the elderly. We will take a deep dive into data from the Veterans Affairs department, and see what it tells us and what else can be done.

## KEYWORDS

i523, hib315, Big Data, TBI, Veterans Affairs, Concussion

## 1 INTRODUCTION

Be it a researcher, a developer, a scientist, a doctor, an accountant, a stay-at-home mom, or a DJ, we all want to know that we are making a difference. For some, it's through one-time or episodic opportunities: service projects for some, for others it comes in the form of making a monetary donation, and for others, it's less formal: simply helping someone in person. But for others, there is an opportunity to know that what they are doing day in and day out could help someone in a meaningful way. It's particularly satisfying that what you may do for a living could indeed help someone in a consequential way. For most of us we wonder if we are fulfilling that yearning within us. The manager in the business that is trying to make a buck, or the employee at the license bureau, or the teaching assistant...those may have to seek out that extra bit of fulfillment or satisfaction.

Yes, indeed, we can all make a difference....even a Data Scientist! Yes, in a small or large way a Data Scientist can make a real difference. Is it the keen Python skills that Data Scientists possess that make the difference? Or is it the machine learning and R-programming that sets the data scientist apart? Not likely, if not, then surely the ability to query a database, provide documentation, and manage a keen bibliography is the critical piece. All of these are important to be clear. However, a Data Scientist can make a true and meaningful difference by knowing and exercising two essential cornerstones of big data: 1) Knowing your data. 2) Being willing to take the road down which the data leads. Sometimes its the journey that determines the destination. Are you willing to go?

## 2 THE BEGINNING

When the author was young, around 10 or 11 years old, he fell off his bike directly in front of his house. He hit his head on the side of a cement sewer and received a concussion. At the time, it was uncommon for a youth to wear a helmet. The idea of wearing a helmet was not even thought about. He was not taken to the hospital or doctor for any treatment or diagnosis. Therefore, there was nothing definitively diagnosed, medically speaking. Nothing was quantitatively documented. No X-Rays or MRI to test or verify. The only tangible *data* was a large knot on the side of his head, a loss of balance so bad that he could hardly walk or even stand upright, and tremendous nausea.

Today there are concussion protocols. At the slightest sign of concussion symptoms, the footballer is taken into a tent and examined by a physician. We have learned the importance of a quick diagnosis and immediate treatment. In today's professional game, we have means of measuring the g-force of each hit, using sensors within the helmet. Wikipedia describes g-force (the "g" referencing gravity) as "a measurement of the type of acceleration that causes a perception in weight. Despite the name, it is incorrect to consider g-force a fundamental force, as "g-force" is a type of acceleration that can be measured with an accelerometer. Since g-force accelerations indirectly produce weight, any g-force can be described as a 'weight per unit mass'. Helmet sensors can indicate the direction of the impact in addition to the g-force measurement. This information is collected and real-time notifications are sent to trainers and even parents of young athletes.

## 3 YOUTH ATHLETICS

In recent years, due to health issues of high profile professional athletes, we are learning more about the long-term health impacts of head injuries sustained as a young adult or as a child. Along with medical research, data has helped to pave the way for a growing understanding of the impacts of TBI (Traumatic Brain Injury) as well as impacts to the head that are not traumatic. This is indeed critical as the incidents occur at a "rate of 1.6-3.8 million in the United States per year with accumulated costs approaching \$60 billion annually. Recent studies have identified relationships between magnitude, frequency and location of these sustained head impacts with post event symptoms and decrements in neurocognitive function." [2]

With mounting evidence that head impacts, as young adults, can impact our long-term neurological health, the previously quoted report from Shockbox research (one of several independent studies) drew still further concern. It concludes, "when monitored for head impacts across a regular season, it was seen that the elementary age players (average age of 13) experience a similar magnitude and frequency of impacts compared to the high school players (average

age of 17). The frequency and magnitude of high peak g impacts in elementary (71 impacts with 50 g average) causing 6 concussions is also similar to the high school team players (84 impacts with 51 g average) who did not have any team concussions. It is also seen that youth players are still developing skills and techniques leading for increased impacts at the front of the head.” [2]

While the “data road” is showing us information in terms of numbers, is the data diverse and comprehensive? That is, we can measure force and impacts to the brain but what about other data sources like biological tests (*markers*) or images? Patrick Bellgowan, a scientist at the University of Tulsa’s Laureate Institute for Brain Research, focused his effort on measurement and *form* of the hippocampus. The hippocampus is a major portion of the human brain and is part of the limbic system and thus plays a role in spacial memory, long and short-term memory, and is linked to processing and likely emotional control. Bellgowan’s research is published in the *Journal of the American Medical Association* and shows some very stark numbers. “A group of 25 players with no history of reported concussions had hippocampuses that were, on average, 14 percent smaller than those of a control group of 25 males of similar age and health who didn’t play contact sports.” [3] Further, the report shows that while “much of the health and safety debate over football and other contact sports focuses on the risk of developing severe, headline-grabbing neurodegenerative diseases like amyotrophic lateral sclerosis (ALS) and CTE, a growing body of evidence suggests that both concussions and subconcussive blows can alter mood, cognition and behavior while causing damage and structural changes to the brain.” [3]

Traveling down this “data road” there is no shortage of information around concussions, TBIs, and athletics. We can certainly understand the concern with the potential long-term impacts on one’s health. Fear for our young athletes’ health would concern any doctor, pediatrician and most certainly parents. Based on the quantity of information one would think that it’s the athlete that is most susceptible. The youth are vulnerable to be sure, but is it only the athlete? It’s clear that professional athletics is big business in our country and around the world. It wasn’t until the deaths of the big name professional athletes that this issue came to the public’s eye. Researchers were already looking into this but focus was not sharpened until we heard the names of Mike Webster, Dave Duereson, Andre Waters and Junior Seau. These athletes made millions and the big business of athletics was at stake. With all this money comes more concern and interest and, in this case, data availability.

However, this “data road” if you will, did not originate with athletics. While the bulk of the data is based on TBI concerns for athletics, it originated with the military. Military personnel have tremendous challenges with head trauma due to the force of explosions and projectiles. With the military, the measured force of explosions and the speed of projectiles is tremendous especially when compared to the trauma experienced by most athletes. Add to this, a soldier is under significant stress as part of his job. How both head trauma along with significant levels of stress impacts the soldier’s neurological health requires much more research, as do related TBIs in the general population due to auto accidents, falls and violence. While culturally the impacts of TBIs on athletes garners greater attention, it’s the military and the population at

large that provides opportunities for the Data Scientist, if we let the data do the driving.

#### 4 THE GENERAL POPULATION

The Center for Disease Control and Prevention (CDC) is a great source for detailed information regarding TBI’s and the general populous. They are involved in providing detailed reports to congress and to the citizens regarding items related to our health. While access to raw data is not available, the breadth and scope of the data that is available is worth studying. The CDC provides data regarding the rates of hospital visits, emergency department visits, and death broken out by age group and type or cause of trauma. Much of this data is available going back several years. In addition, there are numerous reports including broad level statistics. For example:

[Figure 1 about here.]

It’s by knowing the data that you can better understand the quality of the data. If we want to make a difference, in data science, we have to know the source of the data. The CDC leverages a few sources for their TBI related reporting. The Healthcare Cost and Utilization Project (HCUP) is a part of the US Health and Human Services Department. It’s a comprehensive repository or collection of databases related to hospital stays and patient care data. HCUP has yet another collection of databases called NEDS (Nationwide Emergency Department Sample) and NIS (National Impatient Sample). The databases rely on data that is input and managed by hospitals at the time of the patients stay. The CDC feels that the data “sample size is large enough to provide stable annual estimates of TBI.”

One way that these databases are created is due to the efforts at the hospital using the International Classification of Disease (ICD), in this case the ICD-9-CM, which is the *International Classification of Disease, Ninth Edition, Clinical Modification*. To a Data Scientist this is a standardized basis of input of the source data - a very good thing. During a patient’s stay in the hospital, there is a standard in identifying the diagnosis as well as classifications, used for patient releases or secondary diagnoses. For example, lets say a patient is admitted into the hospital with a TBI due to an auto accident in which the patient’s head hits the steering wheel, causing injury. Code 804: multiple fractures involving skull or face with other bones, could be used. Multiple codes could also be used to further define the extent of the injuries or diagnosis. While this approach has inherit challenges for a data scientist, for example, how do we manage analysis tied to multiple codes, or potential administrative issues (loss of data or mis-classification)? The good thing is that we do have a standard and data that we can build upon. And in this case, the data source has a standardization which is important. To help alleviate some of the aforementioned data concerns, we could minimize some of the data challenges by increasing the sample size or extrapolate by leveraging other similar sources to increase our certainty.

The CDC’s published report “Traumatic Brain Injury - Related Emergency Department Visits, Hospitalizations, and Deaths - United States, 2007 and 2013,” concluded the following: “In 2013, approximately 2.8 million TBI-related ED visits, hospitalizations, and deaths occurred in the United States, representing an increase in 2007 that

was largely attributed to an increase in the number and rate of TBI-related ED (Emergency Departments), i.e., Emergency Room visits. Although much public interest has been devoted to sports-related concussion in youth, the findings in this report indicate that older adult falls account for a much larger proportion of the increase in TBI-related ED visits during this period. In addition, although the modest increase in ED visits that might be attributed to youth sports concussions do not extend to increases in TBI-related hospitalizations and deaths, the same cannot be said for TBIs attributed to older adult falls. From 2007 and 2013, increases in TBI-related hospitalizations and death attributable to older adult falls suggest the need for greater attention to preventing older adult falls. Empirically validated prevention measures can help reduce the incidence of older adult falls.” [5]. By quantifying the causes of some of the TBIs for the general public, using data, we could indeed make a difference and help to “empirically validate prevention measures”.

If the ICD codes were expanded to include more details around causation, or if during the hospital visit the symptoms were tagged to *potential* cause we could build upon that from a data perspective. Knowing that some will not know the cause of the fall (elderly), or a patient could have been concussed due to domestic violence and is unwilling to discuss it, building upon what we currently have could help us use the data to help prevent TBIs in the general public. The CDC created the STEADI program (Stopping Elderly Accidents, Deaths and Injuries), which is an effort to help identify older adults at risk and help prevent falls. A Data Scientist research could marry results to those in the medical field for an opportunity to help with predictive analytics and preemptively provide support to the elderly and others at risk using the data we have and are building upon. While there is so much attention paid to athletes and data tied to youth head injury prevention, there is a vast opportunity to help others, of all ages, and truly make a difference.

## 5 THE VETERANS

On the playing field, athletes like to call it the “Field of Battle”. This analog represents the arena in which the athlete challenges himself to beat his opponent and win the game. While the “Field of Battle” for the athlete offers a competitive challenge, it is not life or death. The soldier, fights on THE field of battle. He doesn’t play to win but to live. Balls are not flying but bullets! This places an immeasurable amount of stress on the soldier. Imagine adding a TBI to a critically stressed soldier. The dynamics increase, and so does the need to understand the causes, health, history, and symptoms of the soldier. In short, we need more data in order to help.

“While most people fully recover from a concussion within three to six months, soldiers who suffer concussions in battle can experience symptoms for years following the injury, says Michael K. Rauls, an experimental psychology student at Augusta State University in August. Combat-related stress may prolong soldiers’ recovery, and, at the same time, concussions may hamper soldiers’ ability to recover from stress.”[1]

The following url provided raw data for examination. This is data from the Veterans Affairs database accessible under the catalog of government data. The data sets are available under the catalog section and is intended for public access and use. In addition, Metadata has been created and was updated in November, 2017.

url - <https://catalog.data.gov/dataset/mild-tbi-diagnosis-and-management-strategies> [4]

From the website, once you have accessed the catalog you will notice the source data used is JSON data via the website. JSON (JavaScript Object Notation) is easy to access and is considered “human-readable”. We chose to convert the JSON to CSV (Common Separated Values) then uploaded via Python as well as input into excel. We converted CSV to XLSX, a Microsoft Excel format. We analyzed the data for any obvious issues as this is normally a good time to do some data cleaning. It’s also a good time to do some realigning if need be, i.e., move some data around to align more cleanly. At this point, we used the pivot function within excel. The pivot function is one of excel’s most powerful features. It allows the user to align large data sets in order to extract meaning. The result is a table like format that is easily used to make charts or graphs.

Below is a snippet of the JSON data downloaded from the Veterans Affairs (VA) website.

[Figure 2 about here.]

Using a simple conversion tool, this image is a snippet of the JSON parsed data translated to CSV.

[Figure 3 about here.]

Using Python, via Jupyter Notebook, to pull in the CSV for additional programming work.

[Figure 4 about here.]

There are two data sets: (1) Mild TBI Diagnosis and Management Strategies: Implications for Assessment and Treatment in Veterans (2) Mild TBI Diagnosis and Management Strategies: VA’s TBI Screening and Evaluation Program. The Screening and Evaluation Program document contains primarily data, in terms of total numbers, related to the Veterans symptoms. We will however dig deeper into the Assessment and Treatment in Veterans data.

This data from the Assessment and Treatment dataset is small and straightforward. The data variables have four categories: DataElement, DataType, DataValue and TBIFlag. TBIFlag is used to differentiate those that were diagnosed with TBI as the data set includes both those diagnosed with TBI and those that were not identified. The working assumption is that there could be some that actually did have a Traumatic Brain Injury that may not have been officially diagnosed and flagged in this data. If you looked at it hierarchically, DataElement is the more granular of the two key characteristic variables. So DataElement would be the “child” to the DataType, and is dependant on the DataType.

The specific parsing and analyzing effort was to take the CSV file and import it into Microsoft Excel. Using excel, for both datasets, we created a workbook in which to analyze the data, similar to a Jupyter Notebook used for Python. We created a “data tab” which held the CSV data that we copied and placed into a pivot table. Leveraging the two primary data variables of DataElement and DataType we analyzed the data. Based on what the data told us we set up additional tabs as a work space in the workbook to look at particular DataTypes. At this point, we created charts to illustrate the data findings, the most pertinent charts of which are included in this project. This was done by the charting capabilities within excel.

The DataType includes several categories, but we will focus on a few. One area was the “category of care by patient with and without TBI”. This data set included a total of 684,133 veterans that were patients. Of that, 47,845 (7%) had the TBI flag while the remaining 636,288 did not. Included are some charts created from the data based on the *category of care* and as we expected, those with TBI (flagged for TBI) had a higher percentage of treatment related to Psychiatry and Mental Health.

[Figure 5 about here.]

Additionally, we looked into another DataType category of “Prevalence of Mental Health and Pain”. This category includes DataValues of: depression, bipolar, psychosis, PTSD (Post-Traumatic Stress Disorder) and others. While it was not clear, we are assuming this diagnosis of the prevalence of mental health and pain was made as a result of the treatment of TBI, though the data does include numbers on veterans who had been (prior to) receiving services from the VA (Veterans Affairs/Administration), listed as a “VA User”. The point here is there is some uncertainty as to whether their diagnoses were pre-TBI or were as a result of having a TBI. The results, however, were particularly stark in that the veteran with TBI had a marked increase in areas around mental health, PTSD, headache and depression. The chart shows the comparisons.

[Figure 6 about here.]

One DataType that caught our eye was the “Category of Care Inpatient Length of Stay”, since veterans with TBI and other stress related health issues should have a marked increase in the length of stay. However, the TBI flagged veterans did not show an increase in the time of stay, the data did not represent a length of time. Since time was not a DataElement option, we decided to compare the “In-patient Stay” percentages to the total number of patients and found that it was the result of the two. Therefore, “Category of Care Inpatient Length of Stay”, was not a representation of time but the number of patients. The simple chart illustrates the result. So *specifically* the length of stay is actually the “Number of Patients by Inpatient Care Category” shown in a prior chart. While we feel this is an error with the data, the general feeling is that we think it is minor but will reach out to the VA and advise. However, it would have been interesting for this researcher to see that data.

[Figure 7 about here.]

While we do have errors in the data, the fact that the data does include Diagnosis Codes as a DataType variable is promising. These are the same codes described prior as ICD (International Classification of Disease). This helps to provide us with additional information that can be used to further help medical personnel and support of veterans who have prolonged rehabilitation due to TBIs and the various levels of stress, including PTSD.

The second of the two JSON data sets “VA’s TBI Screening and Evaluation Program”, includes yearly diagnosis numbers and “Post-concussive Symptoms in the last 30 days”. The yearly diagnosis numbers at first blush appear to provide meaningful information as it includes a yearly total of patients as well as the percentage of those that were diagnosed with pain, or PTSD or TBI, which was very similar to the data in the first JSON data set. Additionally, the overall year to year numbers should be reflective of the amount of military activity within that given year, assuming the concussions were directly related to military activity. That is, I would expect

to see the overall numbers, as well as those particular diagnosis numbers, increase when there is military activity. This data does not provide any information regarding the level of activity by the military, within the given year.

However, the second JSON data set includes the “Postconcussive Symptoms in the last 30 days”, which helps to support that data in the first JSON data set “Implications for Assessment and Treatment”. This data set evaluated 55,070 postconcussive veterans within 30 days. My working assumption is that these veterans would have been diagnosed within the last 30 days, but could have sustained the concussion more than 30 days ago. The postconcussive symptoms in this data set are numerous and center around the general categories of: Anxiety, depression, fatigue, forgetfulness, headache, irritability among others. For each symptom category the symptom is measured as either, none, mild or moderate to severe.

The postconcussive symptoms that had the highest number of veterans, that was diagnosed with that symptom, all had moderate to severe measured symptoms. The top five identified symptoms, all having a moderate to severe rating are: 1) Irritability (easily annoyed), 2) Sleep Disturbance, 3) Forgetfulness, 4) Anxious or tense and 5) Headaches. For example, of the 55,070 postconcussive veterans, 45,389 felt irritable and were easily annoyed; which is 82 percent. Over 72 percent complained of moderate to severe headaches. It would be interesting to compare similar raw data from concussed athletes to these numbers to see how they compare.

So what does all this data tell the Data Scientist? The data driven direction has found that veterans and active soldiers with TBI are diagnosed at a higher rate in terms of levels of stress, and thus have a need for mental health treatment and psychiatric care. As a result, recovery takes much longer than say the general population or for athletes and, based on the first JSON dataset, a majority of veterans were Veteran Affair or services users. Given the data that we have, it appears that TBI veterans have as many symptoms and likely many more, and for a longer period of time, than athletes. To that, veterans have the additional complication of stress that increases the symptoms and the severity. Imagine a TBI veteran that is experiencing moderate to severe headaches, anxiety, fatigue and irritability. Not only is the concern around the long-term neurological health, like Alzheimer’s or CTE (Chronic Traumatic Encephalopathy), the veteran has to deal with a prolonged recovery *short-term*.

How can the Data Scientist make a difference? Given the data we have, we can start to create some models to help align the symptoms to care and start to leverage this based on each military mission. In essence, prepare for the TBI patient, coming in from the battle field, regarding long-term care. Since we have the ICD codes (diagnosis codes) in this data, can we use these or enhance the data to draw a connection to the causes of the TBI. If so, by linking the causation, to the symptoms and recovery, we can prepare earlier, plan for the impact and the cost. With the goal of ultimately working with the military to help limit TBIs. The idea would be to use the information to help proactively identify soldiers at risk, diagnose quickly, and be prepared with the necessary treatment as quickly as possible, which includes planning for the future, as unlike the athlete who may need to quit playing a game, the soldier may be giving up his job. Anything the Data Scientist can do to help, would make a difference.

## 6 THE CHALLENGES

The desire of this research was to pull together a fascinating story to show that head trauma, even mild trauma, over a period of time could cause long-term debilitating neurological effects. Furthermore, we wanted to show the benefit of documenting the daily head trauma of football players in both practice and games and then add all the data together and tell the story of how we now know that mild regular head trauma is just as dangerous as TBIs and maybe more so, as you may not know you are in long-term neurological danger. This would have been supported by the output of helmet sensors and collection of data each and every day. The daily data for each player would have been collected, cleaned and organized to show the direction of head impacts (what part of the brain was affected). The data would have included a *g-force measurement* for each hit, again on a daily basis.

All of the aforementioned would be linked to, and compared with, a regular MRI or another type of image. Say for example, an annual MRI of the brain to compare year to year changes or to look for any changes due to the helmet impact data gathered throughout the year. Any potential correlation from these two data sources would be documented. From a data science perspective, we had envisioned a method to represent the images quantitatively so that we could more easily align the image to the daily impact data. For example, we could divide the brain up alphabetically, each letter representing a different area within the brain. To each alphabetically designated section, we could apply a combination of numeric value and a word, say a color, to illustrate both the location, type and severity of injury, or changes to the brain from the MRI. For a Data Scientist then you can start to build some correlation between the daily data and the specific message in the image. Gathered on an annual basis we could build some history and move towards doing some predictive-analytics. To both the daily impact data and the images we could also include biological markers, that could be analyzed from blood to show concussions and other biological changes, further adding *richness* to the data. However, we never found that data. Yes, medically speaking, some correlation or cause-and-effect data exist but not in an overall quantifiable manner and not available in a data set for the Data Scientist.

Finding medical data sets is a challenge to be sure. We researched for days trying to find any data set with little luck. We would expect that medical data sets would be limited due to privacy concerns and how quickly data in the medical world can become stale. We would expect that some data sets were very challenging and expensive to gather and that the owner would not want to freely offer such an asset. Then there is the likely challenge of the complexity of a given medical data set. We would assume that most medical, in particular head trauma based, data sets would be complicated and challenging to access, parse, clean, read, etc. However, it was a learning experience to find so little. It also has us wondering if the medical community might be well served to have some other *eyes* looking at the data. We read several medical studies and journals and know that there is indeed data available based on the research that was found, but nothing that was accessible. As a result, this might be an excellent opportunity for data science, computer science and medical worlds to work together towards a shared goal, and make

a difference. This created a question for us that we were not able to have the time to further investigate.

Since the data drove us to the general public and in particular head injuries to the military, we were not able to pull together information around the causes, other than high level information. That is, by knowing more about the cause of the head trauma we may be able to better work toward limiting them. We simply assumed one thing, but the data steered us in a different direction. It would have been interesting to pursue any roll that big data could have played in helping to limit TBIs in the military. Just the concussion from the delivery of a large artillery piece is significant. Imagine being on the receiving end.

Not only was there limited data, but the JSON data sets were small. The actual patient numbers were reasonable but the lack of historical data based on diagnosis types, or symptoms related to any long-term neurological diseases, or additional data on the patient recovery would have been welcomed. Imagine then comparing this to a similar data set for athletes who had experienced TBI, which would have been very interesting. It's from this point that we could build a cost-estimate related to treatment and recovery and then show the difference in cost for treatment of the athlete and the veteran. For the Data Scientist, working towards pulling together a cost-estimate for veterans would help in preparing the necessary care as well as doing analysis on the benefits of prevention and additional research. As for the veteran, their likely short-term treatment will take much longer than the athlete, not to mention the expense of the veteran not being able to work as well as any potential long-term impacts to the veteran, complexities not all athletes have to endure. In theory, we might find that research dollars are better spent dealing with TBIs in veterans than the athlete.

Another challenge is trying to better understand what we don't know. For example, the JSON data sets were based on those veterans receiving care from the VA. We also know from the data that about 80% of them had been receiving care from the VA prior, but we don't know how far back their VA based treatment and diagnosis went. We also do not know how many soldiers were diagnosed on the battle field or training and are not receiving care from the VA and are therefore not identified in these numbers. To the Data Scientist, it would be important to know of all the military personnel that were diagnosed with TBI, how many are in treatment with the VA?

Lastly, as the paper outlines, we have a concern about the long-term effects of head trauma, especially repeated trauma. Alzheimer's, CTE, Parkinson's diseases are catastrophic, and building an open relationship between these diseases and concussion are ongoing. This is a growing area of research, as the medical research of the brain is complicated and still maturing. Additionally, this type of research is a long-term endeavor as some of these diseases are not diagnosed until years after the head trauma. So the challenge of gathering actionable data can take years - even decades. Also, some of the disease identification requires research of the brain after the death of the patient. It was not until after the doctors were able to examine the brains of athletes that they identified what is now called CTE.

## 7 CONCLUSION

“The road goes ever on and on, down from the door where it began. Now far ahead the road has gone, and I must follow, if I can, pursuing it with eager feet, until it joins some larger way where many paths and errands meet. And whither then? I cannot say.” - J.R.R. Tolkien, The Fellowship of the Ring.

The availability of data drove us down a different road than we expected. It also opened our eyes to a brand new group affected by brain injuries - The Military! While the data sets were limited, the research process placed us on a road that opened our eyes. A road that was built upon not only concern of TBIs for athletes, but for the general population and the military.

When compared to the athlete, the soldier needs so much more to recover. The TBI's are worse, requiring a longer period of time to recover. The extended recover time impacts life in the military and at home, thus impacting the quality of life as well as vocation, as they may be forced to retire or ultimately be disabled. The potential long-term impacts are not insignificant. Their road is one with hills. We can't help but dream of what we can do to help.

With more data, we can make a difference in identifying the injured. With more data we can build towards complete rehabilitation and treatment. And with more data we can help prepare for the financial challenges. With all the resources and companies lined up to help the athlete, as well as substantial financial resources from sports leagues and institutions...we wonder if our veterans will be treated similarly. Will we maintain the quality of helmet sensors for our soldier that the star football player receives? Will we keep a database and track impacts to our soldiers like Riddell helmet company does for the high school football player? If not, the data may just show us what we are not doing for our veterans. That is a road worth traveling one that is making a difference....welcome to Data Science.

## ACKNOWLEDGMENTS

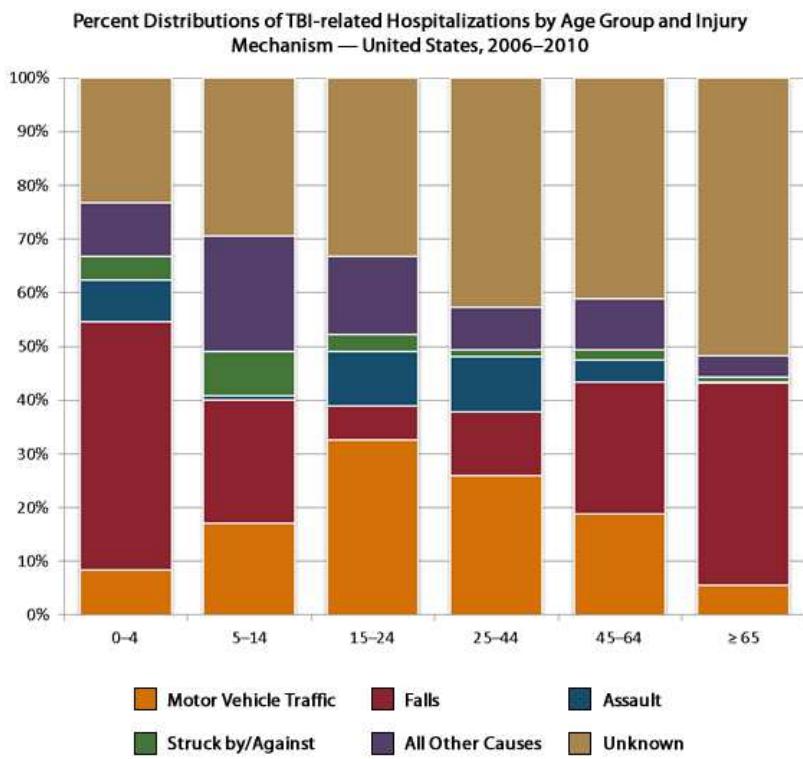
Many thanks to Professor Gregor von Laszewski, the Teaching Assistants and Indiana University. I also want to thank Katie, my understanding wife. Lastly, for my employer AT&T for a commitment to education and giving me 26 years of experience, challenge and opportunity.

## REFERENCES

- [1] Jared Clark. 2010. Concussions spell double trouble for soldiers. Website. (nov 2010). <http://www.apa.org/gradpsych/2010/11/research.aspx>
- [2] D. Crossman and et al. J.E. Bailes MD. 2014. Monitoring of Higher Magnitude Head Impact Exposure in Youth and High School Football Players. 1, 1 (2014), 1–4. <http://www.theshockbox.com/shockbox-research-in-youth-football/>
- [3] Patrick Hruby. 2014. Facing The Truth. Website. (may 2014). <http://www.sportsonearth.com/article/75487104/football-concussions-traumatic-brain-injuries-nfl>
- [4] Department of Veterans Affairs. 2016. Mild TBI Diagnosis and Management Strategies: Implications for Assessment & Treatment in Veterans and VA's TBI Screening and Evaluation Program. <https://doi.org/FederalGovernment>
- [5] Christopher A. Taylor and et al. Jeneita M. Bell. 2017. Traumatic Brain Injury - Related Emergency Department Visits, Hospitalizations, and Death - United States, 2007 and 2013. 66, 9 (march 2017), 1–16. <https://doi.org/10.15585/mmwr.ss6609a1>

#### LIST OF FIGURES

1	CDC TBI Type 2006-2010	8
2	Veterans Affairs JSON	8
3	Veterans Affairs CSV	9
4	Veterans Affairs ipynb	10
5	Veterans Affairs Care Category	11
6	Veterans Affairs mental health and pain	12
7	Veterans Affairs incorrect length of stay	12



	<b>Motor Vehicle Traffic</b>	<b>Falls</b>	<b>Assault</b>	<b>Struck by/Against</b>	<b>All Other Causes</b>	<b>Unknown</b>
<b>0-4</b>	1,116	6,184	1,044	589	1,327	3,123
<b>5-14</b>	2,306	3,077	111	1,118	2,887	3,976
<b>15-24</b>	13,257	2,590	4,131	1,230	5,949	13,517
<b>25-44</b>	15,522	7,045	6,134	777	4,670	25,539
<b>45-64</b>	12,178	15,962	2,668	1,296	6,091	26,775
<b>≥ 65</b>	5,282	36,525	285	912	3,774	50,197

**Figure 1: CDC TBI Type 2006-2010**

```
[{"DataElement": "Patients", "DataType": "Characteristic: Population", "DataValue": "47,845", "TBIFlag": "Y"}, {"DataElement": "Patients", "DataType": "Characteristic: Population", "DataValue": "33", "TBIFlag": "Y"}, {"DataElement": "Characteristic: Population", "DataValue": "13", "TBIFlag": "N"}, {"DataElement": "Characteristic: Population", "DataValue": "636,288", "TBIFlag": "W"}, {"DataElement": "Age Mean", "DataType": "Characteristic: Population", "DataValue": "18", "TBIFlag": "N"}, {"DataElement": "Age Standard Deviation", "DataType": "Characteristic: Population", "DataValue": "8", "TBIFlag": "Y"}, {"DataElement": "Age Standard Deviation", "DataType": "Characteristic: Population", "DataValue": "10", "TBIFlag": "W"}, {"DataElement": "Male", "DataType": "Characteristic: Gender", "DataValue": "94%", "TBIFlag": "Y"}, {"DataElement": "Male", "DataType": "Characteristic: Gender", "DataValue": "86%", "TBIFlag": "N"}, {"DataElement": "White Only", "Characteristic: Race", "DataValue": "75%", "TBIFlag": "Y"}, {"DataElement": "White Only", "Characteristic: Race", "DataValue": "73%", "TBIFlag": "W"}, {"DataElement": "Black Only", "Characteristic: Race", "DataValue": "13%", "TBIFlag": "Y"}, {"DataElement": "Black Only", "Characteristic: Race", "DataValue": "14%", "TBIFlag": "N"}]
```

**Figure 2: Veterans Affairs JSON**

DataElement	DataType	DataValue	TBIFlag
Patients	Characteristic	47,845	Y
Patients	Characteristic	636,288	N
Age Mean	Characteristic	33	Y
Age Mean	Characteristic	36	N
Age Standard	Characteristic	8	Y
Age Standard	Characteristic	10	N
Female	Characteristic	6%	Y
Female	Characteristic	14%	N
Male	Characteristic	94%	Y
Male	Characteristic	86%	N
White Only	Characteristic	75%	Y
White Only	Characteristic	67%	N
Black Only	Characteristic	13%	Y
Black Only	Characteristic	18%	N
Native Amer	Characteristic	1%	Y
Native Amer	Characteristic	1%	N
Asian Only	Characteristic	2%	Y
Asian Only	Characteristic	2%	N
Native Hawa	Characteristic	1%	Y
Native Hawa	Characteristic	1%	N
Multiracial	Characteristic	3%	Y
Multiracial	Characteristic	2%	N
Unknown	Characteristic	6%	Y
Unknown	Characteristic	10%	N

jupyter i523 project cdc data Last Checkpoint: Last Tuesday at 4:29 PM (autosaved)  Logout

File Edit View Insert Cell Kernel Help Trusted Python 3

In [3]:

```
import csv
import pandas as pd
data=pd.read_csv('tbi assessment.csv')

ModuleNotFoundError: Traceback (most recent call last)
<ipython-input-3-3d05bb22dbf3> in <module>()
      1 import csv
----> 2 import pandas as pd
      3 data=pd.read_csv('tbi assessment.csv')

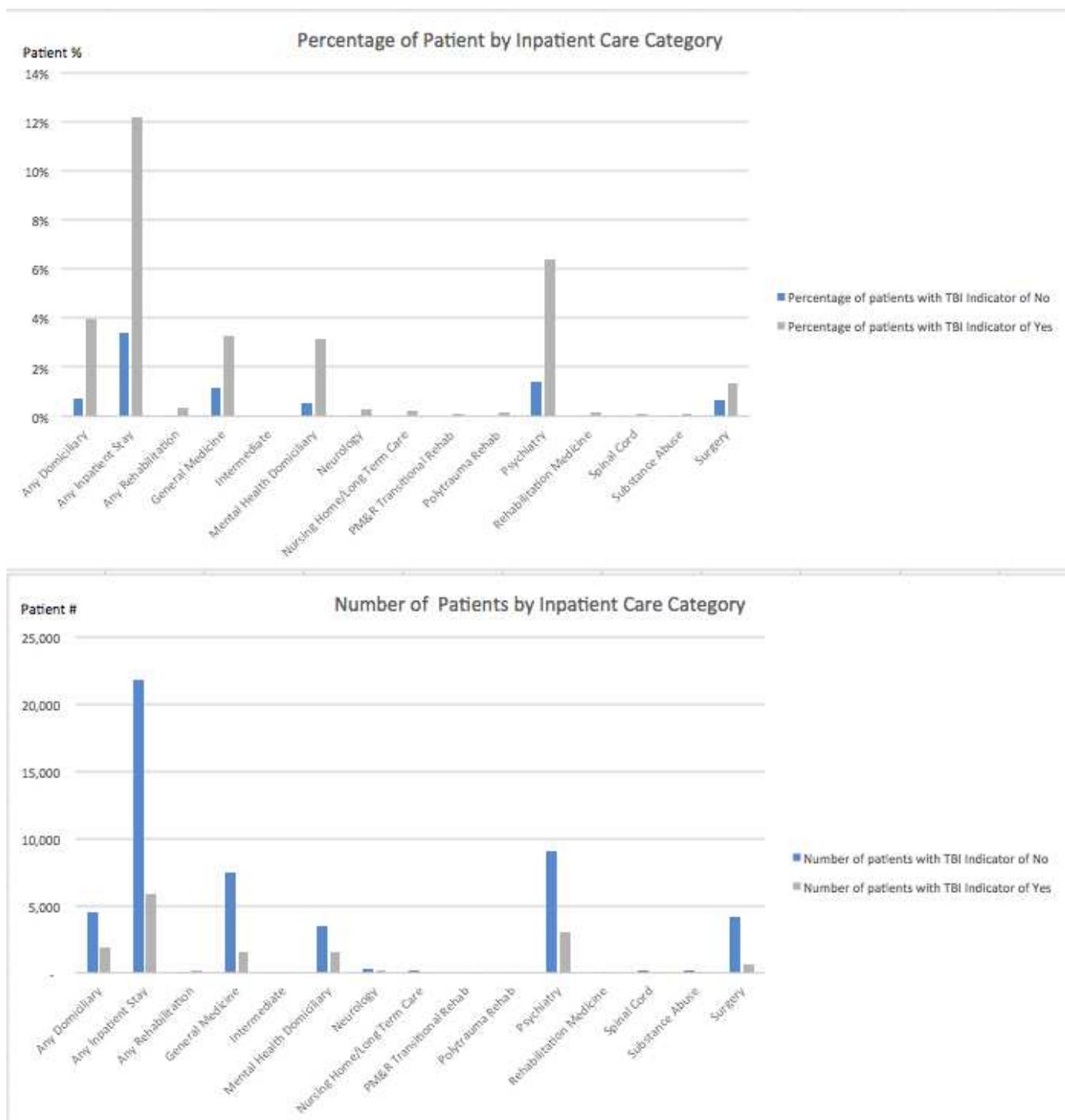
ModuleNotFoundError: No module named 'pandas'
```

In [6]:

```
import csv
with open ('tbi assessment.csv') as file:
    reader = csv.reader(file)
    for row in reader:
        print (row)

['DataElement', 'DataType', 'DataValue', 'TBIFlag']
['Patients', 'Characteristic: Population', '47,845', 'Y']
['Patients', 'Characteristic: Population', '636,288', 'N']
['Age Mean', 'Characteristic: Population', '33', 'Y']
['Age Mean', 'Characteristic: Population', '36', 'N']
['Age Standard Deviation', 'Characteristic: Population', '8', 'Y']
['Age Standard Deviation', 'Characteristic: Population', '10', 'N']
['Female', 'Characteristic: Gender', '68', 'Y']
['Female', 'Characteristic: Gender', '14%', 'N']
['Male', 'Characteristic: Gender', '94%', 'Y']
['Male', 'Characteristic: Gender', '86%', 'N']
['White Only', 'Characteristic: Race', '75%', 'Y']
['White Only', 'Characteristic: Race', '67%', 'N']
['Black Only', 'Characteristic: Race', '13%', 'Y']
['Black Only', 'Characteristic: Race', '18%', 'N']
['Native American/Alaska Native Only', 'Characteristic: Race', '1%', 'Y']
['Native American/Alaska Native Only', 'Characteristic: Race', '1%', 'N']
['Asian Only', 'Characteristic: Race', '2%', 'Y']
['Asian Only', 'Characteristic: Race', '2%', 'N']
['Native Hawaiian/Pacific Islander Only', 'Characteristic: Race', '1%', 'Y']
```

Figure 4: Veterans Affairs ipynb



**Figure 5: Veterans Affairs Care Category**

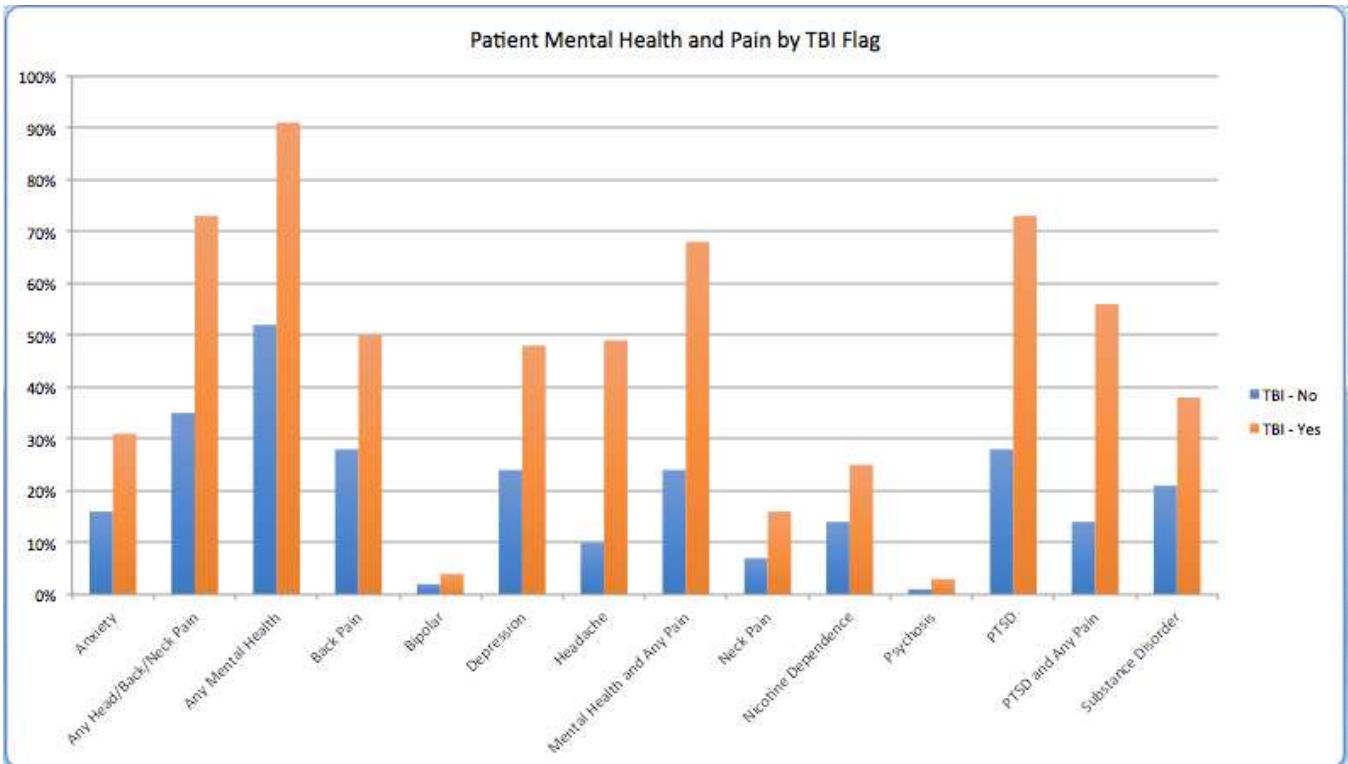


Figure 6: Veterans Affairs mental health and pain

Length of Stay Data				Inpatient Stay % * Total	
Row Labels	N	Y	Row Labels	N	Y
Patience	636,288	47,845	Patience	636,288	47,845
Any Domiciliary	4,540	1,878	Any Domiciliary	1%	4%
Any Inpatient Stay	21,822	5,844	Any Inpatient Stay	3%	12%
Any Rehabilitation	88	170	Any Rehabilitation	0%	0%
General Medicine	7,469	1,573	General Medicine	1%	3%
Intermediate	18	-	Intermediate	0%	0%
Mental Health Domiciliary	3,429	1,514	Mental Health Domiciliary	1%	3%
Neurology	264	140	Neurology	0%	0%
Nursing Home/Long Term Care	203	102	Nursing Home/Long Term C:	0%	0%
PM&R Transitional Rehab	-	44	PM&R Transitional Rehab	0%	0%
Polytrauma Rehab	-	75	Polytrauma Rehab	0%	0%
Psychiatry	9,044	3,074	Psychiatry	1%	6%
Rehabilitation Medicine	76	83	Rehabilitation Medicine	0%	0%
Spinal Cord	131	49	Spinal Cord	0%	0%
Substance Abuse	110	34	Substance Abuse	0%	0%
Surgery	4,147	655	Surgery	1%	1%

Figure 7: Veterans Affairs incorrect length of stay

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no journal in www-theshockbox-com
Warning--empty booktitle in www-catalog-data-gov
Warning--empty publisher in www-catalog-data-gov
Warning--empty address in www-catalog-data-gov
Warning--unrecognized DOI value [Federal Goverment]
Warning--no journal in www-cdc-gov
(There were 6 warnings)
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-12-16 09.36.42] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Float too large for page by 274.2707pt.
'h' float specifier changed to 'ht'.
Typesetting of "report.tex" completed in 1.6s.
```

---

```
Compliance Report
```

---

```
name: Garner, Jeffry
hid: 315
paper1: november 02 2017 100%
paper2: 100% complete
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
12
wc 315 project 12 6003 report.tex
wc 315 project 12 5895 report.pdf
wc 315 project 12 217 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
5: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
66: \begin{figure}[h]
```

```

67: \includegraphics[width=\columnwidth]{images/graph1.png}
68: \caption{CDC TBI Type 2006-2010}\label{f:CDC TBI Type 2006-2010}
95: \begin{figure}[h]
96: \includegraphics[width=\columnwidth]{images/graph2.png}
97: \caption{Veterans Affairs JSON}\label{f:Veterans Affairs JSON}
102: \begin{figure}[h]
103: \includegraphics[width=\columnwidth]{images/graph3.png}
104: \caption{Veterans Affairs CSV}\label{f:Veterans Affairs CSV}
110: \begin{figure}[h]
111: \includegraphics[width=\columnwidth]{images/graph4.png}
112: \caption{Veterans Affairs ipynb}\label{f:Veterans Affairs ipynb}
123: \begin{figure}[h]
124: \includegraphics[width=\columnwidth]{images/graph5.png}
125: \caption{Veterans Affairs Care Category}\label{f:Veterans Affairs Care Category}
130: \begin{figure}[h]
131: \includegraphics[width=\columnwidth]{images/graph6.png}
132: \caption{Veterans Affairs mental health and pain}\label{f:Veterans Affairs mental health and pain}
138: \begin{figure}[h]
139: \includegraphics[width=\columnwidth]{images/graph7.png}
140: \caption{Veterans Affairs incorrect length of stay}\label{f:Veterans Affairs incorrect length of stay}

```

```

figures 7
tables 0
includegraphics 7
labels 7
refs 0
floats 7

```

```

True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are referred to: (refs >= labels)

```

```

Label/ref check
passed: True

```

When using figures use columnwidth  
 $[width=1.0\columnwidth]$   
do not change the number to a smaller fraction

---

```
find textwidth
```

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no journal in www-theshockbox-com
Warning--empty booktitle in www-catalog-data-gov
Warning--empty publisher in www-catalog-data-gov
Warning--empty address in www-catalog-data-gov
Warning--unrecognized DOI value [Federal Goverment]
Warning--no journal in www-cdc-gov
(There were 6 warnings)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
The following tests are optional
```

=====  
Tip: newlines can often be replaced just by an empty line

find newline  
-----

passed: True  
cites should have a space before \cite{} but not before the {

find cite {  
-----

passed: True

# Big Data = Big Bias? An Analysis of Google Search Suggestions

Gabriel Jones

Indiana University Bloomington  
Bloomington, Indiana, USA  
gabejone@indiana.edu

## ABSTRACT

While Big Data can make the world a better place, blind optimism in its infallibility can cause irreversible damage to society if left unchecked. With the mission of ensuring accountability, we debunk the fallacious narratives people tend to tell about Big Data, offering a more realistic discussion of its merits and its limitations. We then explore how analytical or algorithmic bias and sampling bias, two problems that statisticians have faced since long before the onset of Big Data, present pitfalls for deriving knowledge from data. We examine how the ethical implications of these pitfalls can cause serious damage in society. We determine that effective, credible, and ethically sound Big Data analysis must obey the principles of transparency, clear and appropriate objective definition, and self-correcting feedback mechanisms. We examine case studies where academicians and businesses have tested algorithms to study how well they exhibit these principles. We then implement our own test to check for potential algorithmic bias in Google. Based on evidence that certain individuals have been corrupted in part by Google searches allegedly bias against racial minorities, we hypothesize that Google's algorithms systematically exhibit biases against minority groups. We test this hypothesis by examining how Google search suggestions associate certain negative words with names that typically belong to minority groups. We conclude that while our study alone cannot prove or disprove our argument, the evidence in our analysis contradicts our hypotheses, thus suggesting that no systematic bias is exhibited. We discuss end by discussing what the results could mean for future studies of potential algorithmic bias in Google.

## KEYWORDS

i523, hid104, hid216, Big Data, Ethics, Algorithmic Bias, Sample Bias

## 1 INTRODUCTION: FALLACIOUS NARRATIVES ABOUT BIG DATA

Since its origins, Big Data has promised to revolutionize the world. Scholars have wisely noted that it represents a paradigmatic shift from conventional norms of data, but the public has latched onto provocative yet unrealistic narratives that deify Big Data as omniscient, infallible, and impervious to bias. Confiding in such narratives diminishes the integrity of credible science and poses serious ethical challenges, but these challenges are more likely overlooked because the problematic narratives seem to reject the need for ethical discussion.

In 2008, *Wired.com*'s Chris Anderson wrote an article that captures the general optimism with which people conceptualize Big Data. The article, with its self-explanatory title "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", argues that

Mathew Millard

Indiana University Bloomington  
Bloomington, Indiana, USA  
mdmillar@indiana.edu

Big Data provides such a complete, infallible view into reality that we no longer need conventional methods of scientific inquiry but need only to look at what the data tell us. According to Anderson, "With enough data, the numbers speak for themselves"[1]. This fervorous optimism was further extended in a 2013 book by Mayer-Schonberger and Cukier titled *Big Data* where authors assert that Big Data is synonymous with all data. In the past, researchers could only look at samples of data with limited scope, but Big Data, the authors claim, represents not a sample but a complete set[11]. A dataset of Twitter posts is viewed as synonymous with a complete, unbiased set of all of society's thoughts. By analyzing such a dataset, they conclude that they can confidently answer any question about how all of society thinks and behaves[9].

Cheerleaders for Big Data, such as Anderson, Mayer-Schonberger, and Cukier to make five exciting but yet flatly incorrect claims: that bigger is always better; that data analysis produces indisputably accurate results; that every data point can be studied, eliminating the need for archaic statistical sampling techniques; that studying causation is no longer needed since correlational patterns tell us all we need to know; and that scientific and statistical models are obsolete, since Big Data is itself sufficient. They have tended to extrapolate from the early success of the Google Flu Trends which at the time successfully embodied such grandiose, idealistic views. The Google Flu Trends project employed a theory-free set of algorithms that studied search engine results to predict flu outbreaks faster and more accurately than the Center for Disease Control. Allowing the numbers to "speak for themselves", Google determined that the number of searches about the Flu were correlated with flu outbreaks, so they concluded that more searches could accurately predict a greater spread[9].

At first it worked brilliantly. But in February 2013, just a month before the  $n = all$  proposition was published in *Big Data*, it made headlines for failing miserably, overestimating actual trends in 2013 by over 140 percent, leading Google to humbly terminate the program. The overconfidence of such an enormous dataset, viewed as a complete representation of reality free of gaps or inconsistencies, blinded them to its inherent flaws. For one, searches involving the term *influenza* are hardly an unbiased determinant of flu prevalence. They committed a classic statistical mistake by failing to consider confounding variables: the other reasons why people might search for the word *influenza*. Rather than adapting their model to fit changing patterns in the data, they assumed that the numbers could speak for themselves[9].

But blind proponents of Big Data bury the Google Flu Trends fiasco as just one not particularly convincing counterexample, giving superficial explanations that do not challenge Big Data's position as an infallible deity. In reality, such failure is the rule rather than the exception. Even Gartner, a company publicly known for pushing the importance of Big Data, estimated that 60 percent of Big

Data projects would fail[8]. But it's not just a matter of occasional success or failure; many people in all disciplines misunderstand the nature of Big Data and therefore have unrealistic expectations. The narrative of the Target coupon case shows that society still regards the potential of Big Data as omniscient even if its execution is occasionally flawed. The story is narrated somewhat as follows.

In 2012, Target had collected enough purchasing data about pregnant women that they determined a particular high school girl was pregnant. When coupons for baby care items mixed in with general coupons started showing up in the mail, the father angrily visited the store manager to complain, suggesting that the store was encouraging teen pregnancy. The manager understood his frustration and called twice to apologize, but on the second call, the father's mood was different. The father offered his own apology because Target was right. His daughter was pregnant, and Target's Big Data analytics managed to discover this before him[6].

While such a rose-colored narration fits well within the aforementioned grandiose conceptions of Big Data, a closer look shows that this successful case is overblown. While the anecdote seems to prove that Target's algorithms are infallibly accurate – that everyone receiving baby care coupons is pregnant – this is very unlikely. While the popular account suggests that Target mixes in coupons targeted towards pregnant women with other coupons to avoid spooking such women about their algorithmic accuracy, a much more credible explanation is that many women see mixed advertisements precisely because Target is unsure which ones actually are pregnant[9]. Even women who Target does suspect are pregnant have shopping interests outside of baby care items. While the algorithms help not to waste money by sending the coupons to, say, a single male adult living alone, they hardly indicate any reliable accuracy of pregnancy prediction. Of course, this is an empirical question that could be answered by researching how often pregnancy-targeted ads are sent to pregnant women versus those who aren't. But without having a methodologically sound study prove consistent accuracy, it's unwise to extrapolate from the anecdote and assume that Big Data done right is omniscient.

Critiquing the dominant reading of the Target case is not meant to suggest that Big Data has no value. Afterall, Target likely improved the efficiency of targeted advertising through Big Data by more accurately segmenting those who *might* be pregnant. But the important thing to keep in mind is that ultimately, models of the world and the data that feed them are imperfect. Models reflect the biases of those who create them, and data reflect biases inherent in sampling methods, time periods, and society in general. Cathy O'Neal, a former professor and Wall Street algorithm specialist with a mathematics degree from Harvard, observes that any model of the world "begins with a hunch, an instinct about a deeper logic beneath the surface of things"[13]. Human potential for bias and faulty assumptions can creep in. Of course, hunches or working thesis provide a necessary part of the scientific method of inquiry. Human intuition can be useful, as long there exist mechanisms by which those hunches can be evaluated and revised when necessary[13].

Perhaps the most common example of successfully wielding insightful models is depicted by the movie *Moneyball*, based on a true story. Oakland A's General Manager Billy Beane hypothesized that conventional performance metrics were overrated whereas

more obscure measures better predicted overall success. He worked with statistician Bill James to create models that helped Beane decide which players to acquire and which to let go. The once obscure method has become a staple of baseball analytics. According to O'Neal, the model works for three main reasons: it allows for transparent analysis; its objectives are clear and appropriately quantifiable; and it includes a self-correcting feedback mechanism of new inputs and outputs, allowing it to be honed and refined. Models go wrong when they lack these three healthy attributes: "the calculations are opaque; the objectives attempt to quantify that which perhaps should not be; and feedback loops, far from being self-correcting, serve only to reinforce faulty assumptions"[13].

But models are only one factor in determining the efficacy of Big Data analysis. Since the very nature of data analysis is to extrapolate from limited samples, not only must researchers realize that models include human bias, but data itself is imperfect. It's true that data never lie. But it's false to assume they tell the truth. Data by themselves don't say anything; they simply are[4]. No matter how large and complex a dataset, it is always up to researchers to interpret the data to make meaningful claims. This is the essence of the scientific method that some want to reject.

## 2 ALGORITHMIC AND SAMPLE BIAS: THE THREATS THAT NEVER DISAPPEARED

Humans, as imperfect beings, should never assume that our creations are without flaw and bias. In many ways, mistakes and flawed thinking can trickle into the processes we come up with. This is the idea behind the fallibility of models created by humans with respect to algorithms used for handling Big Data. Some algorithms come with biases based on narrow thinking with a broad scope to cover. Other biases come from the assumption that the Big Data set being used is representative of the population when it really isn't. In any scenario, the creator is prone to introducing bias into any given algorithm, which can make it difficult to trust the results that the algorithm produces. With this in mind and considering the importance of specific findings, there is a lot at stake here. In some cases, lives can be changed for better or worse.

Sometimes algorithms, as models laden with the biases of their creators, can unintentionally manipulate readings of data in ways that reinforce false positives. But not all algorithms are wrong. In fact, machine learning shows us that often a well-written algorithm fed with good data can outperform human knowledge on everything from chess to medical diagnosis. But there's a problem with Big Data; it's inherently messy, complex, and distorted. Contrary to popular opinion that views it as a perfect representation of reality – recall the  $n = \text{all}$  proposition – Big Data is a black box where typical issues with data quality hide themselves rather than disappearing. No matter how large or complex the dataset, the old adage still remains true: garbage in, garbage out.

*The Literary Digest* experienced the concept of garbage in, garbage out firsthand during the 1936 US presidential election, which pitted the Republican Alfred Landon against the wildly popular democrat Franklin D. Roosevelt. Roosevelt was particularly popular among the working class, the US majority, whereas Landon resonated well with the upper middle class and elites[9]. *The Literary Digest* Tried to predict the outcome of the election by sending out surveys to its

own subscribers and by looking people up in phone and automobile registries. During the great depression, the people that owned phones, cars, and subscribed to the *The Literary Digest* tended to be more affluent and republican. After sending out 10 million ballots and receiving back nearly a fifth of them, they predicted that Alfred Landon would win with an astonishing 57 percent of the popular vote. They could not have been more wrong. Landon earned less than 40 percent of the popular vote, losing by a landslide[5]. This case has become the archetype example that data from a bias sample will lead to bias results. Increasing the volume of bad data only succeeds in producing a very precise incorrect conclusion, creating a false sense of confidence in something inherently wrong.

Although the *The Literary Digest* used lots of data, by definition their sample did not involve Big Data[11]. But if we reject the  $n = all$  proposition, we can see that Big Data is still a sample and is therefore potentially vulnerable to sample bias. But while any statistically literate person can understand what went wrong with *The Literary Digest*, sample bias with Big Data is much more complicated and difficult to identify. For many people, random samples of social media data appear impervious to sample bias. Researchers conducting Twitter sentiment analyses often claim objectivity in representing the real world accurately, concluding that patterns observed in these vast, complex webs occur the same way offline. Despite the conflation of people and Twitter users, the two are not synonymous. Twitter users are by no means representative of the population. A Pew Research project in 2013 found that US-based Twitter users “were disproportionately young, urban or suburban, and black”[2]. To complicate things further, we cannot assume that Twitter data accurately represent how users behave because users and accounts are not a one-to-one relationship. Some accounts have multiple users, and some users own multiple accounts. Some accounts are just bots that automatically produce content, and some accounts are created and forgotten, going years without use. Furthermore, among active accounts, data are skewed by how some accounts dominate the discourse. Whereas some users post multiple times per day, others use the site only to view content. In fact, 40 percent of active users view content without making contributions, according to 2011 data from Twitter Inc[2]. The notions of what it means to be active, to participate, and to be a user require critical examination that’s almost universally lacking.

The aforementioned examples highlight problems with available Twitter data, but there’s also a problem with the integrity of available data. Twitter only makes a fraction of its data publicly available through its APIs. The supposed firehose of data theoretically contains all public tweets but explicitly excludes data that a user chooses to make private. Furthermore, theory does not match reality as the firehose lacks some publicly available tweets. Very few researchers get adequately full access. Research by Microsoft’s Danah Boyd and Kate Crawford found that rather than a firehose, most have access to a “gardenhose (roughly 10 percent of public tweets), a spritzer (roughly 1 percent of public tweets),” or just select access through whitelist accounts[2]. Not only are protected data excluded, but data samples are not always randomized. So, a more reasonable description of Twitter data would say it takes a skewed sample of the real world population, further skewed by how users and bots create or do not create content, and then it limits the scope of the skewed data in an often opaque, arbitrary manner[2].

Is this data useful? Without a doubt. Is the data so perfect and infallible that we need not concern ourselves with basic principles of statistical and scientific credibility because “the numbers speak for themselves”[1]? Not even close.

If an algorithm could analyze a large, random sample of every word ever thought, spoken, or written by every human throughout their entire life, we could confidently believe that  $n = all$  and make a sentiment analysis that accurately captures how people feel about a certain topic without regard for methods of scientific inquiry; the numbers would “speak for themselves”[1]. But we do not, and probably never will, have that kind of data. Twitter or other social media platforms are no substitute. While understanding the fallibility of Big Data is perhaps not as clear and straightforward as the *Literary Digest* case, society must be responsible by diligently scrutinizing data. To paraphrase loosely from world-renowned consultant Meta S. Brown, the biggest problem with data analysis will always be people failing to admit that data imperfections exist, failing to look for them, and refusing to do anything constructive about the ethical implications of these imperfections[3].

### 3 ETHICAL IMPLICATIONS OF ALGORITHMIC AND SAMPLE BIAS

As we’ve seen, the massive failure of the Google Flu Trends caused embarrassment and wasted Google’s money. But the consequences they faced are relatively trivial, and given the company’s history of learning from the past, they are probably a better company because of the failure. But when Big Data goes awry, the consequences are not always so trivial and localized. Big Data used unwisely has very serious, irreversible impacts upon society. Pervasive overconfidence can make it harder to acknowledge and confront such impacts until too late.

Society’s current failure to address these issues is the topic of Cathy O’Neal’s book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. She argues that these WMDs, referring to Big Data algorithms, have good intentions but often reinforce harmful stereotypes, especially of minorities and the poor, and become opaque models wielding arbitrary punishments. Through her work in the private sector, she has experienced numerous Big Data horror stories, and the book discusses several different failings of Big Data in various contexts.

One common issue associated with Big Data is the notion of self-fulfilling prophecy: the idea that expectations change reality to make it reflect the expectations. If police suspect African Americans to be more likely to commit crimes, they may patrol black neighborhoods more often and proactively hunt criminal activity. Increasing patrols increases the number of arrests, which provides justification to further increase patrols, causing more arrests, and so on. The prophecy that African Americans are more likely to commit crimes becomes adequately reinforced with their higher incarceration rates. But higher likelihood of arrest is not the same thing as being more likely to commit crimes[12].

It should be easy to see how the example of arrest rates is problematic, but somehow incorporating Big Data tends to make people fail to recognize the possibility of self-fulfilling prophecy. In fact, numerous police departments use algorithms that do just this, inadvertently instructing their officers to focus on areas with high

concentrations of blacks. Crime prediction software that attempts to adjust police deployments according to anticipated patterns fail when they confuse more data with better data. Even though they attempt to prioritize violent and serious crime, data generated by relatively insignificant petty crimes, which occur far more often in poor and predominantly minority communities, can overwhelm the system, making it prejudice. Once the petty crime data enters a predictive model, more police deploy into those neighborhoods, and they are more likely to arrest people by their sheer presence and by the perceived threat that those people pose. The increased arrests justify the deployments in the first place[13].

But the danger does not end there. Once people are arrested by these inherently discriminatory processes, Big Data can work to keep them in prison for longer. This is usually not by intention but by flaws in design. Recognizing how unconscious bias can affect sentencing decisions, courts in 24 US states have started to use computerized models to help assess the risk of recidivism, the likelihood of repeat offense. The models attempt to use Big Data to avoid a common, serious problem with human reasoning, and they certainly show some promise in this regard. But over reliance on the models can prove even worse than trusting potentially biased judges. “By attempting to quantify and nail down with precision what are at root messy human realities”, the recidivism models shroud sentencing bias in a veil of unwarranted confidence and precise accuracy that disadvantages minorities by subjecting them to harsher prison sentences[13].

How does one quantify something as complex as the risk of recidivism? One popular model uses a lengthy questionnaire that attempts to pinpoint factors related to this risk. The questionnaire inquires about things such as previous police incidents. Given how much more often young black males get stopped by the police, partly because of the aforementioned self-fulfilling prophecy, such questions easily become a proxy for race, despite intentions to reduce this very prejudice. Other questions, such as whether or not the respondent’s relatives or friends have criminal records, would be flagrant violations of court procedures and surely elicit objections from a defense attorneys if raised during a trial. But the opacity “of these complicated risk models shields them from proper scrutiny”[13]. Discriminatory police strategies feed into the recidivism models used to call for harsher sentencing, creating “a destructive and pernicious feedback loop”[13].

It is no secret that racial tension has become a dominant source of discussion when it comes to the American justice system. However, this issue is compounded with bias produced within the data itself as well. When there is a bias in how arrests are made based on the color of someone’s skin, this bias feeds into an algorithm which opens up for more bias down the road. As more people of a given color are arrested and given harsher sentences, this data builds up in the system. The root of the cause may be human bias, but there is definitely a healthy amount of algorithmic bias that compounds and builds on the issue as most algorithms lack the ability to look beyond the face value of the data provided[7].

Big Data is, of course, not only used in attempts to more effectively dole out punishments. Facing international competition, Corporate America has latched onto its potential for increasing profits through more effective marketing, financial trading, and personnel decisions. With the prevalence of the internet, social

media, and information literacy, Big Data presents an enormous opportunity for market personalizing. Rather than targeting advertisement campaigns on broad, general audiences, Big Data can segment down to the individual level, targeting people based on their own personal data and patterns of behavior. However, this type of marketing is still a very inexact science and raises tricky ethical issues, including gender bias. Like racial bias, gender bias comes about in scenarios where profiling usually happens. For instance, advertising on the internet aims to reach its intended audience in order for businesses to sell products and make profits. Big Data and the statistical analysis involved might suggest that a certain gender has specific tendencies or lean on embedded societal stereotypes which cause some serious bias in an algorithm. One example might be a job opportunity being advertised. In this case, we want to say that either gender should be shown the advertisement a near equal amount, but we know from experience and outrage that this is not the case. It is almost staggering how it would favor the male population at times, especially when dealing with high paying jobs. Here, we also have a combination of Big Data and algorithmic bias working hand in hand to create biased results that ultimately lead to insult and faulty representation[3].

Beyond marketing, Big Data has found particular popularity among Wall Street investment firms, and for good reason. The ability to incorporate Big Data into decision making has tremendous potential for profitability. But the subprime mortgage crisis demonstrated how this can also have tremendous destructive potential. Financial models exhibited a particular bias, reinforcing the idea that what has worked in the past or what works currently will continue working indefinitely. But the sophisticated mathematical models lacked self-correcting feedback that could indicate inherent flaws. Since the models were driven by the market, if they led to maximum profits, they were considered infallible. Otherwise, why would the omniscient invisible hand of the market reward it? In hindsight we all recognize that betting on the subprime mortgage bubble was a losing proposition, yet the myopic reliance on the market proved disastrous in 2008. During the financial crisis, the algorithms used to assess securities risk became smoke screens. Their complex, mathematically intimidating design “camouflaged the true level of risk”[13]. The opaque models also lacked a healthy feedback mechanism that could have identified the problem[13]. The severity of the 2008 recession shows that companies are not only accountable for their own success and failure. Their misuse of Big Data had broad sweeping effects across the entire economy.

Perhaps it is reasonable to understand why companies might get carried away in a practice that, at least on the surface level, does not appear to affect humans directly. A trader working on the top floor of a Wall Street skyscraper might not see how the work of his mathematicians might hurt or harm average people. But Big Data also plays a role in ways that very clearly affect individuals, especially with the increasing popularity of integrating technology into personnel decisions. Since personnel decisions directly impact company performance, workforce management has become popular, particularly programs that promise to eliminate the guesswork from hiring by screening potential employees [13]. Many of these programs use personality tests to try and automate the hiring process; 60 percent to 70 percent of prospective employers, according to Deloitte Consulting.

Despite the optimism, such tests face the same problem as the recidivism surveys: they try unsuccessfully to quantify and precisely measure “what are at root messy human realities”[13] The high use of personality tests goes against research that consistently shows them to be poor predictors of future job performance. They don’t provide this goal but rather an illusion of objectivity and simplicity. They generate raw data that get plugged into efficient algorithms and give clear answers, as opposed to the time consuming and obviously subjective process of human interviewing. Not only does this illusion coolly deceive companies, it leaves prospective employees disgruntled and confused by results from a opaque systems. Rejected employees don’t know if they’ve been flagged or what caused them to be. The personality tests also lack important feedback mechanisms. There is no way to identify inherent errors in the model and use those mistakes to refine the system[13]. Far too often, personality tests fail both the companies that use them and the prospective employees that get arbitrarily denied a chance.

In each of these cases, the story repeats itself where ethical issues that are normally fairly obvious become invisible when Big Data enters the picture. The argument is not that we should reject the positive potential of a reality that will only grow stronger with time. Rather, we should remain cognizant that a failure to adhere to basic principles of scientific credibility and ethical reasoning can affect people in unseen but deadly ways.

#### 4 POTENTIAL ALGORITHMIC BIAS IN GOOGLE: THE DYLANN ROOF CASE

Sometimes, algorithmic bias can morph and distort opinions in ways that almost seem like indoctrination in nature. In some cases, it can seem like this bias can be the root of a terrible downward spiral into blatant racism, but when do we justifiably point blame at the machine rather than a person’s inner desires? In today’s society, it can be tempting to take the easy way out of tough situations and place the blame anywhere else that might make sense as long as it provides some kind of vindication. That being said, we do live in a generation that is gradually becoming more influenced by the internet and technology in general as the years fly by. With that in mind, it is reasonable to see where a flaw or bias in an algorithm can have a monumental impact in a negative way on some people. Unfortunately, there have been cases where people are significantly effected by these algorithmic biases in ways that trigger a violent disposition towards another group or race.

In 2015, a man named Dylann Roof shot and killed nine people at a church in Charleston, South Carolina. The interesting details hanging around this massacre to make it stand out were the people he shot and the line of reasoning he used to explain how he was eventually led to commit such an act. The attack was done on what was reported to be a predominantly black church which led people to label the offense as a hate crime. However, Roof’s explanation on what might have led him to that point is what makes this story stand out from other hate crimes. In an article that the National public radio published titled “What Happened When Dylann Roof Asked Google For Information About Race?” it was reported that Roof’s defense had made a case that there was more to the act than just simple racism and white supremacy[10]. The argument that the internet had a direct influence on what Roof believed and

that he was acting on the information he was being fed through other sources was being made. Roof elaborated on the subject and explained that it had all begun with the growing popularity of the Trayvon Martin case. Trayvon Martin was an unarmed black teenager who was shot and killed in 2012. After researching the details of the case and coming to his own conclusions he states in a quote in the article “this prompted me to type in the words ‘black on White crime’ into Google, and I have never been the same since that day”[10]. The article continues to dive deeper into what Roof might have encountered. Anyone who has encountered a search engine in general has been faced with the auto complete feature that provides calculated, popular options to give the user some direction. In this case, the potential algorithmic bias surrounding the racial tensions might have led Roof down the path of searching for examples of crime committed by people of color on white people. The National Public Radio itself reported that they tested out Google’s search engine by typing out the beginning of the phrase Roof mentioned and they were prompted with the auto complete option of the exact phrase before they could even type in the word white[10]. Even today, you can perform the same experiment and come up with the same results.

Unfortunately, the main factor in driving this algorithmic bias is popularity and relevance which are hard variables that are difficult to counter and account for in most cases. This means the objective of removing the type of algorithmic bias that Dylann Roof encountered would be difficult and require a major change in how search suggestions and results are calculated. However, this needs to be discussed and changes need to be made or more people will continue to be influenced negatively by algorithmic bias which would put more lives at risk down the road. After all, Roof was only seventeen when he began down the line of thinking that led him to commit those murders. There are numerous children and young adults that have unlimited access to the internet and are wide open to the same influence. So, preventative measures need to be taken in order to assure that we do not see similar stories surface. In order to get that done, there has to be some kind of analysis of bias within specific algorithms to understand them and create ways to account for this bias. If unchecked, not much can be done, but the knowledge from understanding the potential bias in these algorithms could prove invaluable.

#### 5 CASE STUDIES IN CHECKING FOR ALGORITHMIC BIAS

Before we can test for algorithmic bias, we have to have a grasp or understanding on how typical algorithms implemented in concepts such as prediction work. In many fields where predictive work is being done in research, there is a common use of natural language processing algorithms such as Support Vector Machines, Neural Networks, and Naive Bayes. Each algorithm uses a specific type of processing that makes it stand out from the others, but the different aspects to each come with advantages and drawbacks that make each one valuable in certain circumstances. However, it is the internal structure of what makes up the algorithm that allows for bias to get in.

## 5.1 Support Vector Machines

A Support Vector Machine is a supervised learning model that is known for the analysis of data for classification and regression. This algorithm is much better at classifying problems than normal logistical regression, and it always converges to a global minima when classifying data. However, this algorithm can be complex especially when the data isn't linearly separable. In the case that the data set being used is non-separable, the data can be transformed using a kernel function, such as a log function in some cases, in order for the algorithm to fit the data and classify correctly. The main bias that we can see from this algorithm is in the case of it classifying data to the point where patterns form based on things like popularity and relevance which could lead to something similar to what we discussed in the case involving Dylann Roof. Since Support Vector Machines are supervised learning models, any bias can be accounted for in the manual make up of the model because it does need some direction. In that case, bias might come in the form of human error though.

## 5.2 Neural Networks

Neural Networks are algorithmic models in which weighted inputs are fed into a function to compute an output. In a predictive setting, Neural Networks are fairly flexible in structure which allows it to be applied in many scenarios. In this type of model, we can expect to utilize logic gates in order to understand how certain scenarios must be weighted to account for a specific outcome or result that we want. If an outcome relying on three variables hinges on a specific variable being untrue and requires at least one of the other variables to be true, we would assign a higher weight to the first variable than the other two because that one variable to ruin the outcome. We can see how this type of modeling can be used to model complex predictive problems and could possibly be used in a search engine if desired. Based on input strings that make up common and popular words and phrases, we could make an auto complete function using neural networks to put weights on certain inputs to predict the possible resulting desired search for the user. Unfortunately, placing weight on certain inputs and pieces of the function can inherit some bias along with it. Even if those weights aren't manually decided, popularity and relevance can have a large impact on calculation of weights which would still lead to some bias down the road. In the search engine example, this would lead to people being influenced to search something based on the beginning of their search query.

## 5.3 Naive Bayes

The Naive Bayes classifier is a probabilistic model that applies Bayes' Theorem that makes predictions based on a set of training data consisting of predetermined features. However, Naive Bayes places the assumption on the features that each one is independent of the other. Although this stipulation does well enough in real word situations, this assumption can hold it back from working in more complex situations where features are highly dependent on one another. In the world we live in today, we rarely encounter situations in large data sets, especially in Big Data settings, where the features involved are independent. This strong assumption of independence here is where the source of a decent amount of

bias could come from when using this algorithm. Because of this, there is a limitation to how well one can make predictions in more complex situations. After all, Naive Bayes is a probabilistic model, but probabilities tend to rely on a multitude of varying factors which may include the influences of other features involved. With this in mind, we can see how the results may be skewed or flawed, but how exactly does that introduce bias? Other biases can come within and there are undoubtedly some in this algorithm, but another major source of bias can come after the work is done. A critical piece comes from the analytic interpretation of the results. If the probabilities are flawed, then the conclusions made from the results could be flawed as well. In this case, we would call it analytical bias. Even though this type of bias isn't an internal problem of the algorithm, it is a byproduct of it in the end result.

## 5.4 How to Check and Account For Algorithmic Bias

After getting to know what kind of algorithms we are working with and what kind of biases can arise from them, our attention should naturally gravitate towards ways we can check and account for these biases. In some of these algorithms, we can recognize and pinpoint potential biases based on how it acts and the conditions set, but we can also account for a lot of it since Support Vector Machines and Neural Networks are, for the most part, supervised models. This means that some of the biases stem from some programming error which can be accounted for by fixing and tweaking the algorithm. This works for these models because there are main points such as variable weighting and classification in Neural Networks and Support Vector Machines respectively. However, the Naive Bayes classifier is a little more ambiguous. This algorithm struggles with over fitting the data which means that it caters too closely to the data set and prevents it from effectively predicting and adding future data points. In this case, more bias or alteration to the model may be needed to find an equilibrium and allow proper analysis. This requires that less influence be put on handpicking features and allow for a hands off approach. This is often considered soft feature selection in data analytics and is a way of removing the harshness of imposing desired features. With all of this in mind, it is near impossible to completely remove the influence of bias in any given algorithm. There will usually be a flaw somewhere no matter how hard we try to create a perfect model. This is clear when looking back on the Dylann Roof case study about Google algorithmic bias. In the same article, Google indicates that they are aware of biases in their algorithm and that they have worked on eliminating and fixing the issues[10]. Even with this awareness and implementation of fixes, we need data and analysis to help us evaluate how well Google improved their algorithm. This was one of our many motivations for testing Google's search suggestions for bias.

## 6 OUR CASE STUDY: TESTING GOOGLE FOR NEGATIVE SEARCH SUGGESTIONS BIAS AGAINST CERTAIN RACES

Based on the observed need for scrutinizing algorithms and inspiration from researchers who have successfully done so, we devised an experiment to test Google's search suggestions for potential

bias against names associated with particular races. In the following sections, we discuss our methodology, our hypotheses our algorithm, and our results.

## 6.1 Methodology

Building off of several researchers and the anecdotal record which have identified cases of bias against certain names, we decided to test for such biases by identifying how negative word associations correlate with searches for particular names. We decided to look for instances of the words associated with *arrest*, *murder*, *crime*, *homicide*, and *prison*, including the words *arrested*, *arresting*, *arrests*, *murderer*, *murderers*, *murdering*, *crimes*, *criminal*, *criminals* *criminality*, *homicides*, *homicidal*, *prisoner*, and *prisons*.

Since we expected these associations to carry bias based on race and ethnicity, we decided to use data that separated names into racial and ethnic categories. To make as objective of an assumption as possible regarding which names were associated with which race, we used data from the 2010 Census that identifies what percentage of people self-identify with which types of names. This allowed us to classify a name as belonging to a certain racial or ethnic group if most people with that name identified under that race. However, one complication of using Census data is that we only had access to surnames. Whereas, based on intuition, a person might be able to distinguish a predominantly Black first name from a predominantly white first name, this is much more difficult for surnames. However, whereas black and white surnames are less distinguishable, Asian and Latino or Hispanic names are much easier to identify because they are usually not based on the English language. For these names, we can be confident that Google recognizes the name and can make associations based partly on these racial and ethnic identities.

Given that the central argument from our qualitative analysis of Big Data suggests that Big Data analysis often has inherent flaws, we determined that avoiding many of the common pitfalls observed with such research should be a top priority. We wanted to study something with which we could be completely confident that the results reflect the true nature of Google's algorithms, whether they support our hypotheses or not. For this reason, despite initially considering it, we decided to avoid studying web page data for such information. Data from web pages would have been problematic because it is difficult to infer the context behind word associations. Simple word counts could create potential false positives. For instance, if a web page from the search *Mueller arrest* uses the word *arrested* n amount of times, it could be an arrest record report for someone with the last name Mueller, a decidedly negative view, or it could a page talking about the people arrested by order of former FBI director James Mueller. Since our study depends on isolating not only words but negative associations, this lack of context introduces an irreconcilably high level of uncertainty: it would be far too difficult to determine which sites held negative views of the particular name and which ones had positive views. For this reason, we decided instead to look at search suggestions. Google's algorithms play the central role in search suggestions, and the simplicity of the searches largely eliminates the problem of insufficient context. By just typing in a name and the first few letters of a search term, we can be more confident that whichever

suggestions are revealed reflect clearly distinguishable sentiments (negative or positive) regarding the name itself.

## 6.2 Hypotheses

Given the anecdotal evidence and from a history of racial and ethnic bias in the United States, we expected to see a bias toward and ethnic minorities. However, we expected that words associated with criminality would apply much more to Blacks and Latinos than to Asians. Therefore, we hypothesized the following:

*H1: Surnames that most often identify Black people will, on average, have more negative word associations than names which most often identify White people.*

*H2: Surnames that most often identify Black people will, on average, have more negative word associations than names which most often identify Hispanic or Latino people.*

*H3: Surnames that most often identify Hispanic or Latino people will, on average, have more negative word associations than names which most often identify White people.*

*H4: Surnames that most often identify Hispanic or Latino people will, on average, have more negative word associations than names which most often identify Asian and Pacific people.*

*H5: Surnames that most often identify White people will, on average, have more negative word associations than names which most often identify Asian and Pacific Island people.*

## 6.3 Algorithm

The coding for our analysis was conducted with Python 3.5.2. The necessary import modules include: *requests*, *json*, *csv*, *urllib.request*, *matplotlib*, *pandas*, *scipy*, *numpy*, and *timeit*, all of which we installed using *pip install*. We gathered the data for our project from the website for the 2010 Census. Their .csv provided us with a list of 162,254 surnames. We read in the file, which is posted in our github account, using *csv.DictReader*. We isolate the name and race/ethnicity columns. For each name, the race/ethnicity categories include the percentage of people that identify as part of a particular race or ethnicity. Because increasingly obscure names are likely less identifiable, we decided to use only the first 500 names. None of these 500 names were predominantly multiracial or predominantly Native American, so these categories were excluded from our analysis. We created a dictionary with the analyzable categories, *pctwhite*, *pctblack*, *pctapi*, *pcthispanic* and sorted the 500 names into the appropriate categories using a series of *for loops* that checked for the highest percentage.

We then move to the main goal of the experiment: capturing Google search suggestions. Using Mozilla Firefox 5.0 as a user agent, we use a url that automatically outputs a two-dimensional list with the search term and ten Google suggestions. We then define a function to modify the url for each search term and for

every name. When, for a given name, a search suggestion contains one of the target words described in the methodology section (arrest, crime, homicide, etc.), the function adds a point to the name's value. We run the function five times for each target word, creating five dictionaries that store names as keys and the number of negative word associations as values. We add the five dictionaries to a list that we can iterate through. This section involves a lot of searching, so we use `timeit` to make sure it runs efficiently. It typically takes about 200 seconds to run with 8 gigabytes of ram and an Intel i7 processor with 500 names, so we estimate that this step would take about 65,000 seconds or over 18 hours to run with the full data.

Now that the search results are appropriately organized by name and by search term, we need to calculate the scores for each race/ethnicity category. We create two dictionaries for each of the four race/ethnicity categories, with search terms as keys and values defaulted to empty lists. For each pairing, one will be later converted to averages and the other will stay as raw data. We then define a function that will appropriately fill the values so that we know how many negative word associations go along with each search term for each race. As inputs, the function takes the name of the racial/ethnic category, the name of the dictionary for that category (from the four we just created), and the list containing the five search term dictionaries we created earlier. The function uses three for loops to iterate through the various dictionaries and lists to match names with their racial categories and add the values to the scores accordingly. We run this function eight times in total for the four pairs of dictionaries.

To calculate the aggregate word associations for each racial category, all we need to do is sum the values of each key for each dictionary. However, to be completely sure that the function for the previous section worked properly, we implement a set of for loops to calculate the aggregates instead of just summing the dictionary values. We create a dictionary with keys set to the racial category and values defaulted to zero and allow the for loops to calculate the proper scores. We also create a dictionary with racial categories as keys and empty lists as values. We append the numerical values for each word association to the empty list. With the data arranged as such in raw form, we are now able to analyze more measures of central tendency, including the median and the 1st and 3rd quartiles. This data arrangement enables us to create the box plot displayed in the results section. Now, for every pair of category dictionaries, we use one of the dictionaries to calculate averages of the aggregate data so that we know how many negative word associations go along with each name and each search term on average.

As a final data preparation step, we are now ready to conduct statistical analysis through a separate .csv file. We write out the data name-by-name to a .csv file by using a series of *for loops* that arrange the data row-by-row grouped by the race it belongs to. The *for loops* arrange everything into a long string with new line characters to separate where each row should be, and then we write it out in one step. From there we read in the new `resultsdata.csv` file with `csv.DictReader` and sort the data totals data into numpy arrays by racial/ethnic category. We then conduct two sample t-tests to compare each array to calculate statistical significance. With this step, we now have all the data summarized and organized as needed and can move on to creating a bar chart, a box plot, and a series of radar charts using `matplotlib`. Details on how to recreate these

visuals are available on the Jupyter notebook on our github project folder.

## 6.4 Results

While we had qualitative data that suggested Google's algorithms could be biased against Black and Hispanic/Latino minorities, the results do not support most of our hypotheses. Regarding H1, while Figure 1 and Figure 2 appear to show that Black surnames do have more negative associations than White surnames, the results were not statistically significant at the .95 confidence level ( $p > .85$ ), so we conclude the null hypothesis in this case. This lack of statistical significance is likely due to the very small number of Black surnames (9) compared to White names (361) in the top 500.

[Figure 1 about here.]

[Figure 2 about here.]

Comparing Blacks to Hispanics with H2, however, Figure 1 and Figure 2 do clearly support our conclusion. The difference in means is about 2.4, and this difference is significant at the .95 confidence level ( $p < .002$ ). With a mean of just 2.9, the average negative associations for Hispanic/Latino was actually the lowest in the data set, so H3 was not just nullified, but the exact opposite is true with statistical significance: White names had more associations than Hispanic/Latino names. Hispanic/Latino names also had fewer negative than Asian/Pacific Island names, so we also reject H4. While Asian/Pacific Island names did not have the lowest amount of negative associations, the group did have significantly fewer than both White and Black names, so we can confirm H5. H5 and H2 were the only hypotheses that were supported by the data; H1 was nullified, and for both H2 and H3, the opposite was shown to be true. Given that our hypotheses were not supported, we cannot support our central claim that Google search suggestions would have inherent biases against Black and Hispanics/Latino surnames based on negative word associations dealing with the identified search terms.

Along with our other visualizations, we included a few radar charts showing the distribution of negative search results based on the word we focused on, as seen in figure 3 and figure 4. One with the four races separated into their own chart and the other with all four races overlapping in one radar chart. The idea behind this observation was to get an idea of which words impacted the negative results the most. The charts do not have a significant impact on our original hypotheses, but are there to just give us general insight into the data. However, by observing the charts, we can see that the three most negatively associated search words across all races were crime, arrest, and prison in no particular order across all races.

[Figure 3 about here.]

[Figure 4 about here.]

## 7 DISCUSSION

While our results did not support our assumptions, this does not necessarily mean that Google does not exhibit any forms of algorithmic bias. However, it does suggest that, Google is probably quite aware of accusations of bias, like the ones following the Dylan Roof case, and has taken active steps to minimize such bias. It could

also mean that just analyzing surnames instead of first names is too generic for Google to recognize racial differences, or perhaps that our search terms did not accurately capture common negative word associations that people relate to certain races. In either case, these provocative results establish the need for more complex algorithmic bias analysis to continue holding such algorithms accountable and rewarding ones that effectively minimize bias.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski for providing the opportunity to explore a topic of deep interest.

## REFERENCES

- [1] Chris Anderson. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Website. (June 2008). <https://www.wired.com/2008/06/pb-theory/>
- [2] Danah Boyd and Kate Crawford. 2011. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. In *Six Provocations for Big Data*. <https://ssrn.com/abstract=1926431>
- [3] Meta Brown. 2017. Math Isn't Biased, But Big Data Is. (AUG 2017). <https://www.forbes.com/sites/metabrown/2017/08/30/math-isnt-biased-but-big-data-is/#2d6691dd4d56>
- [4] Kate Crawford. 2013. The Hidden Biases in Big Data. (April 2013). <https://hbr.org/2013/04/the-hidden-biases-in-big-data>
- [5] Cynthia Crossen. 2006. Fiasco in 1936 Survey Brought 'Science' To Election Polling. (Oct. 2006). <https://www.wsj.com/articles/SB115974322285279370>
- [6] Charles Duhigg. 2012. How Companies Learn Your Secrets. (Feb. 2012). <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?r=1&hp=&pagewanted=all>
- [7] Laurel Eckhouse. 2017. Big data may be reinforcing racial bias in the criminal justice system. (FEB 2017). [https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d\\_story.html?utm\\_term=.0ee1409ec5c0#comments](https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d_story.html?utm_term=.0ee1409ec5c0#comments)
- [8] Laurence Goasduff. 2015. Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. (Sept. 2015). <https://www.gartner.com/newsroom/id/3130017>
- [9] Tim Harford. 2014. Big data: are we making a big mistake? (March 2014). <https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0>
- [10] Rebecca Hersher. 2017. What Happened When Dylan Roof Asked Google For Information About Race? *National Public Radio* (JAN 2017). <https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylan-roof-asked-google-for-information-about-race>
- [11] Carl Lagoze. 2014. Big Data, data integrity, and the fracturing of the control zone. *Big Data and Society* 1, 2 (NO 2014), 1–11. <https://doi.org/10.1177/2053951714558281>
- [12] Jasmine Liu. 2017. Big data and the creation of a self-fulfilling prophecy. (April 2017). <https://www.stanforddaily.com/2017/04/05/big-data-and-the-creation-of-a-self-fulfilling-prophecy/>
- [13] Wharton. 2016. 'Rogue Algorithms' and the Dark Side of Big Data. (Sept. 2016). <http://knowledge.wharton.upenn.edu/article/rogue-algorithms-dark-side-big-data/>

#### LIST OF FIGURES

1	Bar chart comparing average word associations by race.	11
2	Box plots showing distribution of average word association by race.	12
3	Series of radar charts showing average word association by race and by search term.	13
4	Consolidation of the previous four radar charts into one image that shows how each race compares.	14

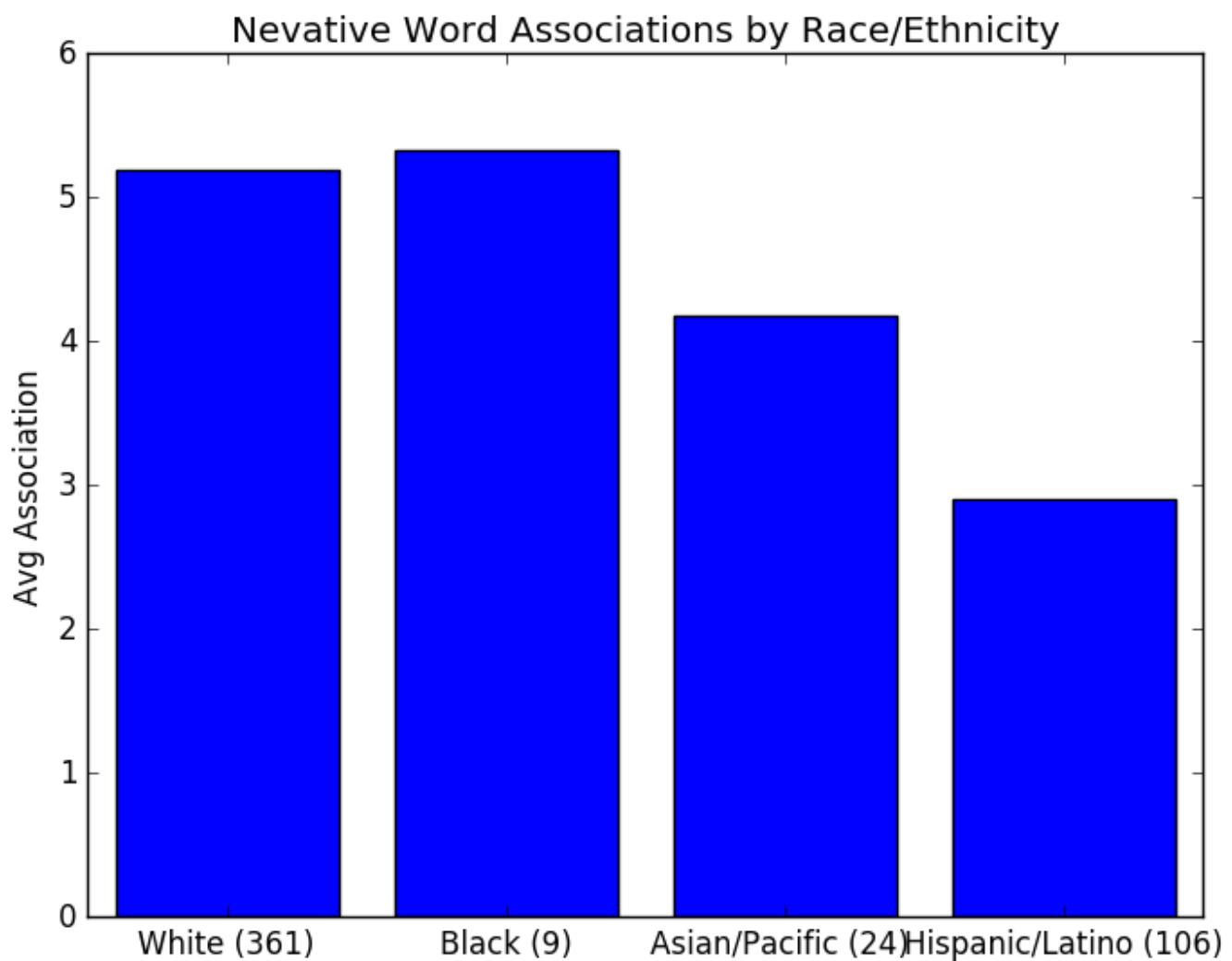


Figure 1: Bar chart comparing average word associations by race.

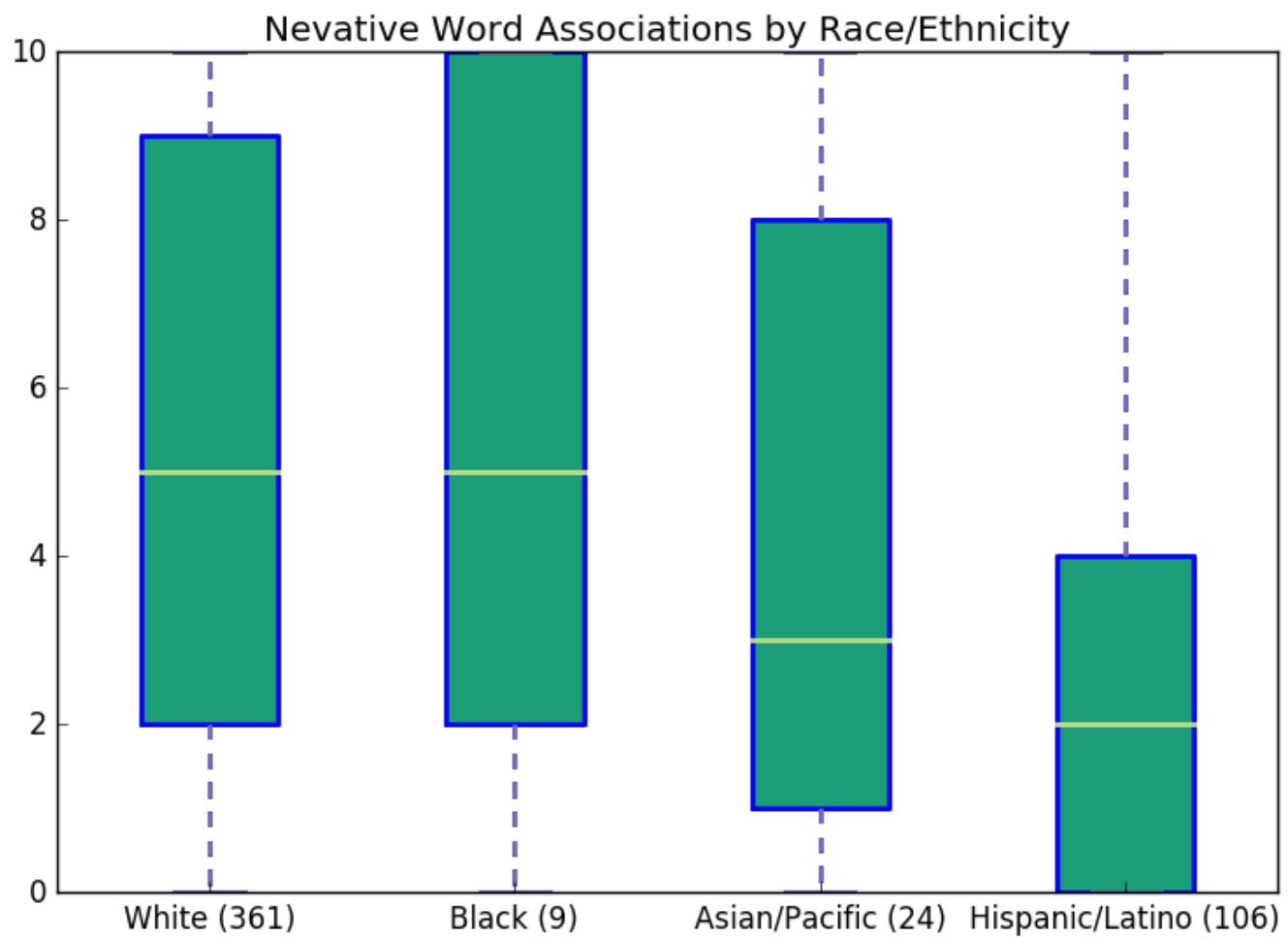
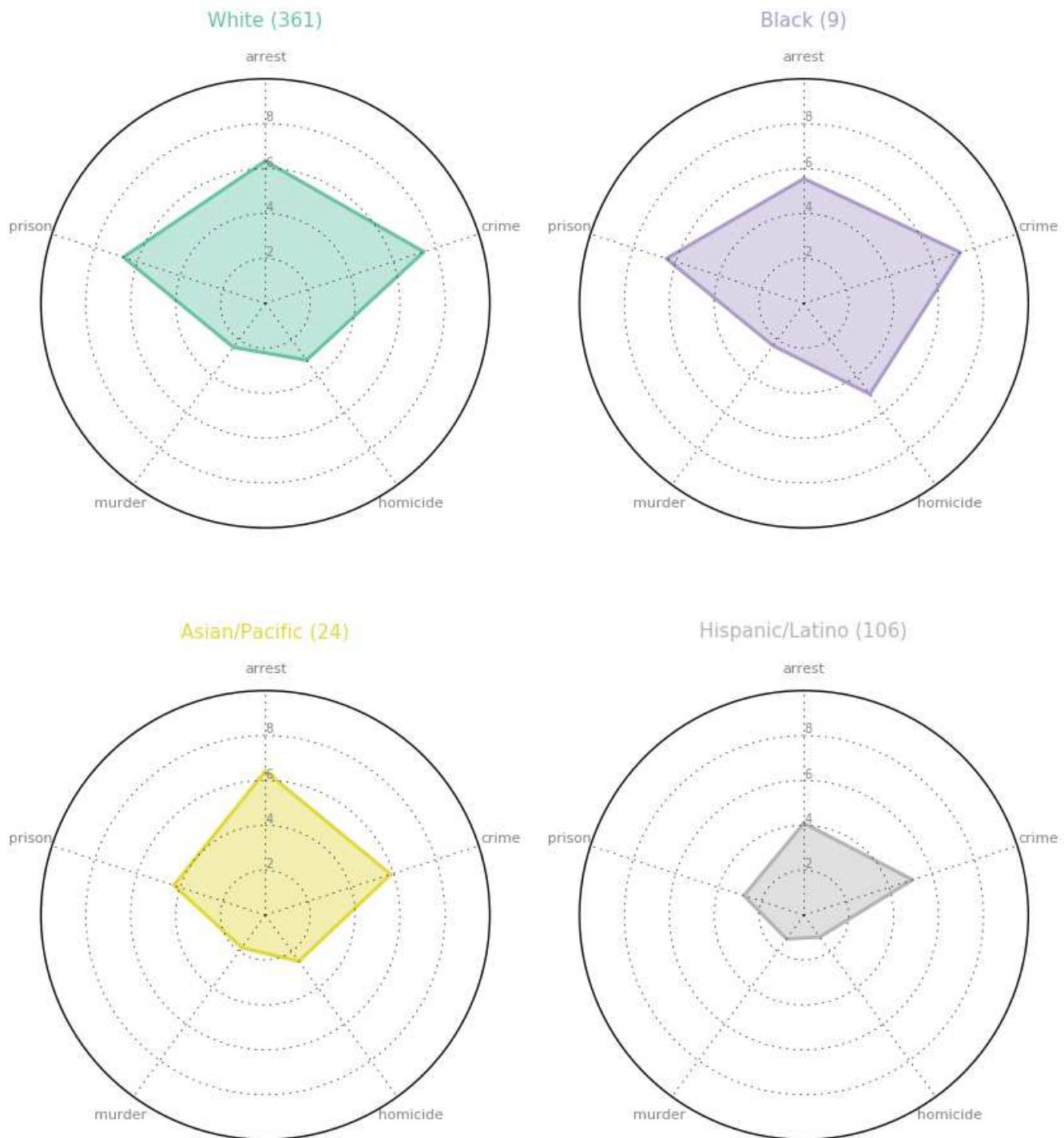


Figure 2: Box plots showing distribution of average word association by race.



**Figure 3:** Series of radar charts showing average word association by race and by search term.

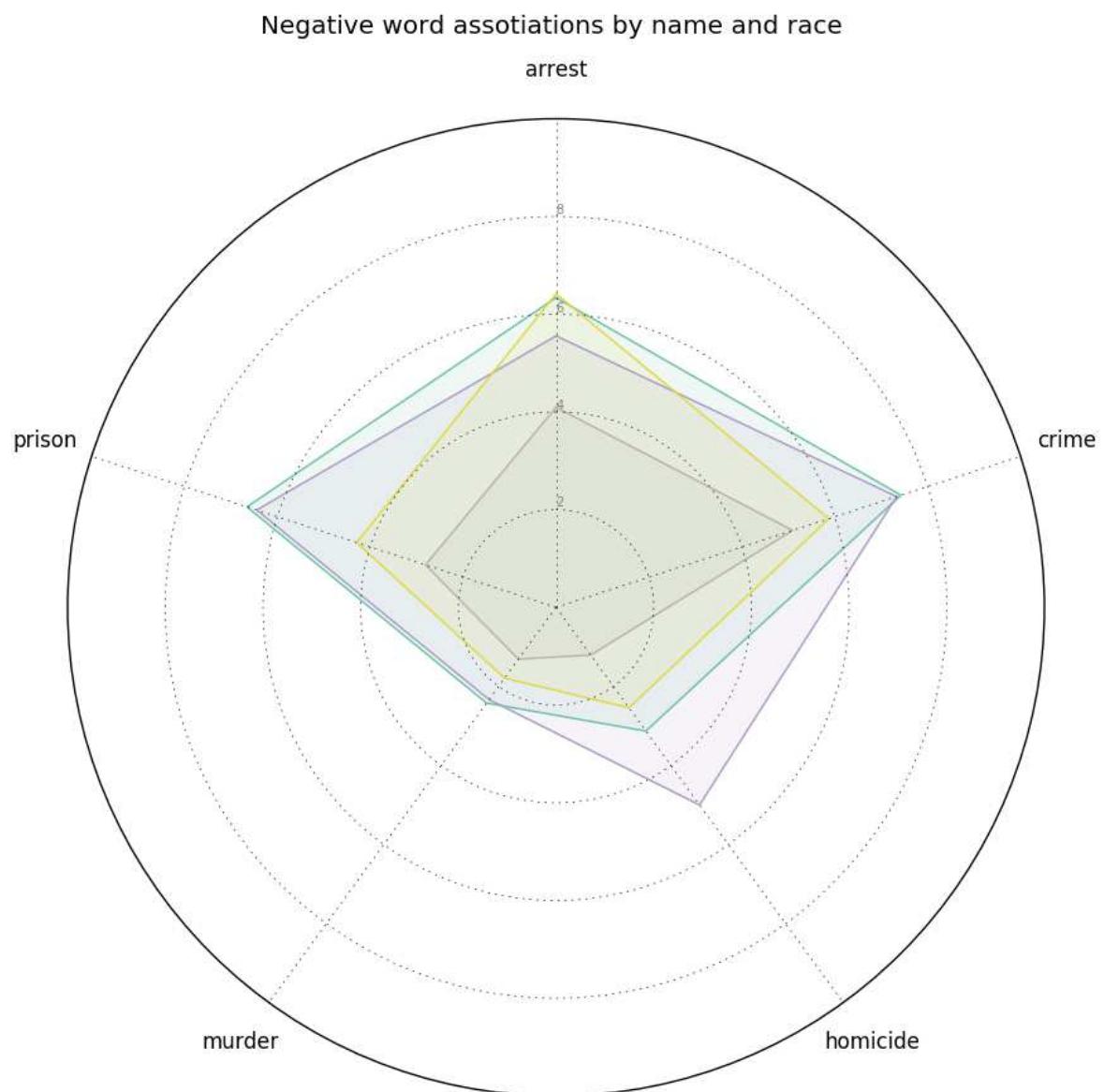


Figure 4: Consolidation of the previous four radar charts into one image that shows how each race compares.

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in Boyd-Crawford2011
Warning--empty address in Boyd-Crawford2011
Warning--page numbers missing in both pages and numpages fields in Boyd-Crawford2011
Warning--no number and no volume in Hersher2017
Warning--page numbers missing in both pages and numpages fields in Hersher2017
(There were 5 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.31.43] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Font shape 'OML/LinuxLibertineT-TLF/m/n' undefined using 'OML/nxlmi/m/it' instead for sys
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Some font shapes were not available, defaults substituted.
Typesetting of "report.tex" completed in 1.4s.
./README.yml
64:73     error    trailing spaces  (trailing-spaces)
65:73     error    trailing spaces  (trailing-spaces)
66:72     error    trailing spaces  (trailing-spaces)
67:77     error    trailing spaces  (trailing-spaces)
68:67     error    trailing spaces  (trailing-spaces)
69:76     error    trailing spaces  (trailing-spaces)
70:77     error    trailing spaces  (trailing-spaces)
71:78     error    trailing spaces  (trailing-spaces)
```

```
72:77    error    trailing spaces  (trailing-spaces)
73:74    error    trailing spaces  (trailing-spaces)
74:70    error    trailing spaces  (trailing-spaces)
75:68    error    trailing spaces  (trailing-spaces)
76:77    error    trailing spaces  (trailing-spaces)
77:81    error    line too long (81 > 80 characters)  (line-length)
78:81    error    line too long (82 > 80 characters)  (line-length)
78:82    error    trailing spaces  (trailing-spaces)
79:79    error    trailing spaces  (trailing-spaces)
80:81    error    line too long (85 > 80 characters)  (line-length)
80:85    error    trailing spaces  (trailing-spaces)
81:80    error    trailing spaces  (trailing-spaces)
82:81    error    line too long (87 > 80 characters)  (line-length)
82:87    error    trailing spaces  (trailing-spaces)
83:81    error    line too long (84 > 80 characters)  (line-length)
83:84    error    trailing spaces  (trailing-spaces)
84:81    error    line too long (84 > 80 characters)  (line-length)
84:84    error    trailing spaces  (trailing-spaces)
85:81    error    line too long (83 > 80 characters)  (line-length)
85:83    error    trailing spaces  (trailing-spaces)
86:81    error    line too long (85 > 80 characters)  (line-length)
86:85    error    trailing spaces  (trailing-spaces)
87:81    error    line too long (83 > 80 characters)  (line-length)
87:83    error    trailing spaces  (trailing-spaces)
100:1    error    too many blank lines (1 > 0)  (empty-lines)
```

---

## Compliance Report

---

```
name: Jones, Gabriel
hid: 104
paper1: Oct 28 17 100%
paper2: Nov 10 17 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
```

```
wc 104 project (null) 8899 report.tex
wc 104 project (null) 8780 report.pdf
wc 104 project (null) 465 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

```
find input{format/i523}
```

---

```
5: \input{format/i523}
```

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
179: \begin{figure}
180: \includegraphics[width=\columnwidth]{images/fig1.png}
182: \label{Figure 1}
185: \begin{figure}
186: \includegraphics[width=\columnwidth]{images/fig2.png}
188: \label{Figure 2}
195: \begin{figure}
196: \includegraphics[width=\columnwidth]{images/fig3.png}
198: \label{Figure 3}
201: \begin{figure}
202: \includegraphics[width=\columnwidth]{images/fig4.png}
204: \label{Figure 4}
```

```
figures 4
```

```
tables 0
```

```
includegraphics 4
```

```
labels 4
```

refs 0  
floats 4

True : ref check passed: (refs >= figures + tables)  
True : label check passed: (refs >= figures + tables)  
True : include graphics passed: (figures >= includegraphics)  
False : check if all figures are referred to: (refs >= labels)

#### Label/ref check

- 176: While we had qualitative data that suggested Google's algorithms could be biased against Black and Hispanic/Latino minorities, the results do not support most of our hypotheses. Regarding H1, while Figure 1 and Figure 2 appear to show that Black surnames do have more negative associations than White surnames, the results were not statistically significant at the .95 confidence level ( $p \text{ \textgreater} .85$ ), so we conclude the null hypothesis in this case. This lack of statistical significance is likely due to the very small number of Black surnames (9) compared to White names (361) in the top 500.
- 176: While we had qualitative data that suggested Google's algorithms could be biased against Black and Hispanic/Latino minorities, the results do not support most of our hypotheses. Regarding H1, while Figure 1 and Figure 2 appear to show that Black surnames do have more negative associations than White surnames, the results were not statistically significant at the .95 confidence level ( $p \text{ \textgreater} .85$ ), so we conclude the null hypothesis in this case. This lack of statistical significance is likely due to the very small number of Black surnames (9) compared to White names (361) in the top 500.
- 181: \label{Figure 1}
- 187: \label{Figure 2}
- 190: Comparing Blacks to Hispanics with H2, however, Figure 1 and Figure 2 do clearly support our conclusion. The difference in means is about 2.4, and this difference is significant at the .95 confidence level ( $p \text{ \textless} .002$ ). With a mean of just 2.9, the average negative associations for Hispanic/Latino was actually the lowest in the data set, so H3 was not just nullified, but the exact opposite is true with statistical significance: White names had more associations than Hispanic/Latino names. Hispanic/Latino names also had fewer negative than Asian/Pacific Island names, so we also reject H4. While Asian/Pacific Island names did not have the lowest amount of negative associations, the group did have significantly fewer than both White and Black names, so we can confirm H5. H5 and H2 were the only hypotheses that were supported by the data; H1 was nullified, and for both H2 and H3, the opposite was shown to be true. Given that our hypotheses were

not supported, we cannot support our central claim that Google search suggestions would have inherent biases against Black and Hispanics/Latino surnames based on negative word associations dealing with the identified search terms.

- 190: Comparing Blacks to Hispanics with H2, however, Figure 1 and Figure 2 do clearly support our conclusion. The difference in means is about 2.4, and this difference is significant at the .95 confidence level ( $p \text{ \textless\textgreater} .002$ ). With a mean of just 2.9, the average negative associations for Hispanic/Latino was actually the lowest in the data set, so H3 was not just nullified, but the exact opposite is true with statistical significance: White names had more associations than Hispanic/Latino names. Hispanic/Latino names also had fewer negative than Asian/Pacific Island names, so we also reject H4. While Asian/Pacific Island names did not have the lowest amount of negative associations, the group did have significantly fewer than both White and Black names, so we can confirm H5. H5 and H2 were the only hypotheses that were supported by the data; H1 was nullified, and for both H2 and H3, the opposite was shown to be true. Given that our hypotheses were not supported, we cannot support our central claim that Google search suggestions would have inherent biases against Black and Hispanics/Latino surnames based on negative word associations dealing with the identified search terms.
- 192: Along with our other visualizations, we included a few radar charts showing the distribution of negative search results based on the word we focused on, as seen in figure 3 and figure 4. One with the four races separated into their own chart and the other with all four races overlapping in one radar chart. The idea behind this observation was to get an idea of which words impacted the negative results the most. The charts do not have a significant impact on our original hypotheses, but are there to just give us general insight into the data. However, by observing the charts, we can see that the three most negatively associated search words across all races were crime, arrest, and prison in no particular order across all races.
- 192: Along with our other visualizations, we included a few radar charts showing the distribution of negative search results based on the word we focused on, as seen in figure 3 and figure 4. One with the four races separated into their own chart and the other with all four races overlapping in one radar chart. The idea behind this observation was to get an idea of which words impacted the negative results the most. The charts do not have a significant impact on our original hypotheses, but are there to just give us general insight into the data. However, by observing the charts, we can see that the three most negatively associated search words across all races were crime, arrest, and prison in

no particular order across all races.  
197: \label{Figure 3}  
203: \label{Figure 4}  
passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

passed: True

---

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--empty publisher in Boyd-Crawford2011  
Warning--empty address in Boyd-Crawford2011  
Warning--page numbers missing in both pages and numpages fields in Boyd-Crawford2011  
Warning--no number and no volume in Hersher2017  
Warning--page numbers missing in both pages and numpages fields in Hersher2017  
(There were 5 warnings)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
```

---

```
The following tests are optional
```

---

```
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# **Big Data Analytics in Support Filtering Wrong Informations On Social Networking Sites**

Juan Ni

Bloomington, Indiana 47401

nijuaniu.edu

## **ABSTRACT**

In an era of information, people are more likely to get information from the ciber world. Due to the conflict of interest, many organizations hire “Spammers” to post a mass of wrong comments under some famous person’s post on Social Networking Sites for control the trend of public opinion[2]. So when user want to see the public opinion under some famous person’s post, they usually get the wrong information which doesn’t represent the real public opinion. Big Data analytic can provide information filtering to screening the fake comment based on data mining technique, and let the user be able to see the true of public opinion on social networking sites.

## **KEYWORDS**

523, HID 107, project, big data, weibo, spammers, data visualizationThChina

## **1 INTRODUCTION**

In highly informatization modern society, internet especially for social networking website carried mass information data. Along with the growth in users at social networking websites, social networking website become a platform which content infinite potential business opportunities and interests. Famous social networking website like Facebook and twitter already became leader which can drive the public opinion direction. I think the influence by social networking websites is completely different than traditional social media, social networking websites are more emphasize on audience’s acceptance and follow suit. For example, some politician announce some idea on news paper and TV news, the influence from it only work when it can make arouse sympathy to the audience, which mean it only work when the audience think it make sense for them. But social networking websites are difference, Seiter mention that“ Comments are a powerful emotional driver. Make the most of them by engaging often with your Facebook community and replying to fans’ comments to keep the conversation going.”[? ]. Social netorking website can make the user believed their opinion by drive their user follow the crowd because people love to view the comments and post comment, “A previous study showed that 45% of users on a social networking site readily click on links posted by their fifriendfi accounts, even if they do not know that person in real life” [? ]. For example, the President of United states Trump really like post his opinion via twitter, we can see lot of people post comments under his tweets, and also many people retweet his tweets. According to Seiter’s statistic, “to let others know what I believe in and who I really am (37%)” [2] is place on the fourth position at social website seeking primarily ranking. This draw a conclusion that if people saw retweets or comments from some Twitter who they believe that twitter is believable, people

will believe the retweets and comments from that twitter is making sense for them. Furthermore, the power of comments under the hot tweets is really powerful because making comment make user no longer be a spectator, they are actually involve into the event and be part of the society. Then, the comments with most retweets will gain people’s trust, and making people think that comment is represent the main strain. So this is how social networking websites impact the main social opinion trends.

## **2 THE ADVERSE EFFECT FROM SPAMMERS**

The social opinion trend at social networking community will drive personal and even company decision, then some trends might harm someone’s benefit because the power of social opinion trend is so powerful. Then people hired spammers to spread wrong information that lead the trend become advantage for them, but the users are become victim because they will make wrong decision because of the trend is control by someone on purpose. “Brown showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks.” [5], every spammer post must for some reason that beneficial for their employer, the most famous case for spammers the shampoo case at 2010. ”BaWang shampoo” is the most famous shampoo at China which advertised by super star Jackie Chan, ”Next Magazine” post a fake news claimed that using ”BaWang shampoo” could cause cancer [3]. I clear remember at that time, almost all social websites post new claim ”BaWang shampoo” is harmful at the same time without any authority judgment, and they put this new at the headline position to abstract user’s eye-ball. Even the authority department proof this new is unreliable, the business reputation of ”BaWang shampoo” had been damaged, lot of people around me stop using this shampoo any more. This case seems have no spammers involved, but actually the spammers for this case is social networking websites themselves instead of single person. The reason why they post slander is because they can get benefit from other shampoo companies in china, other shampoo companies can have more sales because the market-share of ”BaWang shampoo” will be decrease at this case.

The other reason why spammers getting so popular at social networking website is the operating cost of spammers is supper low. Try searching ”buy Facebook like” at Google, and there are over hundred million results come up. And the price of buying like and followers from that website is pretty low, so spammers can get an account with 1000 followers which look like a real account for only 5 dollar [6]. So spammer can create thousands this kind of fake account for posting the wrong information at social networking website, and the detected system is hard to find out those kind of account is real account or Zombie account those accounts have actually followers and like, even feeds. Then spammers can

use script to post batch of wrong information via those account. Furthermore, the cost for post batch of wrong information is unbelievable low; according to the internal information which provided from my friend who working at a IT company, there two ways to post wrong information at social networking website, first one is money reward system, second one is posting AI. For the reward system, a professional spammer company usually have about 10 teams, each team has 500 people; They use reward instead of constant salary, each feed related to the order topic is worth 0.5 Chinese dollar(equal to 5 cent in dollar), and each comment on the target feed is worth 0.2 Chinese dollar( equal to 3 cent in dollar), and the price of long text post is negotiated. The spammers accounts are provide from the company, the price for those account is also low; Most social networking websites only require email address for registered, then they buy email accounts from the retail like 100 Chinese dollar(equal to 15 dollar) for ten thousand accounts, and using script to register social websites accounts and making follow with each fake accounts. This is how they operate the spammer, now most social networking websites register require phone number verify, then they working with local sim card retail which infinity phone number on hand, but the price for each fake account is increase a lot like 3 Chinese dollar(like 50cent) per one, but still pretty cheap compare with other advertise way. Cheap labor force and the development of script technology rising the spammer company, but the lower the user experience at social websites because lot of trash information full of the social websites, people are hard to see the true at social websites any more. According to the network sites worldwide ranking[7], we can see WeChat has almost twice active user than Weibo, and WeChat is the most popular social networking app in China because spammer can not do any at that app. In WeChat, they post feed at the module which call "Friends circle", the feed formatting is pretty similar as weibo, but the app is semi-closed which mean it is complete private, user can only see their friend's post and the comment, no retweet allowed, and if someone who not in the user friend's list comment at user's feeds, user can not see that guys comments. Plenty of users quit traditional social networking website, and the semi-closed social networking app is getting popular, so the adverse effect of spammer is not only for the spam target but also for the platform. If we let spammer keep development, and don't have any way to filter their information, the traditional social websites will die soon.

### 3 BACKGROUND

According to the worldwide statistics data, "Sina Weibo" has 368 millions active users which more than 328 millions of twitter active user [? ], so I would like to using "Sina Weibo" as my investigate target instead of using Twitter. The other reason why I choose "Sina Weibo" as my investigate target is because I'm familiar with Chinese culture and I have been using "Sina Weibo" for more than 8 years. I think my knowledge about "Sina Weibo" will help me a lot at this project and better understand how spammers works at "weibo". The page frame at "weibo" is pretty similar to Twitter [figure1].

[Figure 1 about here.]

The four buttons under each feed are "collect, retweet, reply, like", and the capability for each button is same as twitter. When

user click into the "rely" button, user can see all the comments related to the current feed, and sort them by the amount of "like" that comments get from other user. The only things differen at "weibo" is user not only can see the comments but also can see the retweet information, twitter only allow user to see who retweet the feed. Then user can see the retweet's comments and sort the list by the retweet's times of the retweet feed. So people would love to check the retweet list to see which famous person retweet the feed, and what comment they put into the retweet. Spammer control the public opinion trends by putting wrong information that doesn't represent real public opinion into the comment for some hot feed, they utilize user's habit to reach their goal.

### 4 RELATED WORK

There are many researcher done previous research about how to distinguish the authenticity of information that post on social networking websites. Kr point out the user's social networking structure and the user's feed can represent the credible of the user, and kr using different order algorithm to rank the credible order based on the user's social networking structure and user's feed, and use it to judge the user whether is spammer or not [7]. According to Liu's idea, personal information source is really important for judge the source is reliable or not, like the user register time, the user operating frequency, and the relationship between the user and comment target will be three factors for supervise fake account [17]. In lou's article, he mention that we need to analysis the content and feed for judge the reliable level of information, he also point out the if only investigated the comment, retweet for detect spammer, it is hard to reach automatically fast and accuracy result, which mean it still require operator to control the analysis application [15]. Xu has really unique investigate area, she investigate the spammer in online business platform, and her idea can be work on social networking website [14]. In her article, she focusing on the speciality of spammer's behavior in online business websites, she collect sixty thousand comments and thirteen thousand product information that related to those comments at Amazon, she use those data to analysis the characteristic of user behaviour and set up a classifier for different characteristic of user behaviour; she also use the relationship between different spammers to improve the level of accuracy for detecting spammer.

### 5 METHOD

The method for Filtering spammer's Information at social networking websites can be divided into two part, first part is collecting data and the second part is produce data. Collecting data is the main part at this project because any analysis must base on the data, if the application can not collect the target data from third party platform, then is no way to start analysis.

#### 5.1 Data Collection

Using python 2.7 to collecting data from Weibo, and using the official SDK as my accessed method. First, setting a feed as the investigated target, I using [https://weibo.com/5305999252/Fy0sio7nQ?from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment](https://weibo.com/5305999252/Fy0sio7nQ?from=page_1005055305999252_profile&wvr=6&mod=weibotime&type=comment) this feed to investigated the comment content. The person send this feed is my favorite gaming live streaming player, his name

is LuBen Wei. He is the most popular gaming live streaming in China, there are over four million audiences what his playing game every night. Moreover, this is his second account, so he always post some feeds that can not be post at his official account at Weibo, but there are still thirty thousand comments under this feed , the number of comments at this feed even more than the comments under every signaler feed from Donald J. Trump's account. And the feed's content is he complain about the cheating case, he announced that he never cheating at "PlayerUnknown's Battlegrounds", he claim that the rumor about his cheating is come from the spammers. After he post this feed, this feed became the top 1 hot feed at the feed ranking at Webo, and most comments under this feed are abuse him cheating. So I though there are must be spammer working under this feed, the comments at a feed from a gaming live streaming player's second account is more than the comments from United states's president's feed which is so ridiculous. Therefor I think there must be spammer involved into this feed, that is how we pick up the feed which involved spammer in social networking website. If a feed has unusual comments and likes amount compare with other feed post from the owner, that feed have huge possibility that involved spammer work.

First of all, Weibo require we use Weibo API with authentication, so we need to create a personal application first at the weibo application apply page [13]. Then the weibo official suggest us to use SDK to access the the API, so I came to the sdk websites [8] to get the Weibo SDK package. Normally can just type "pip install sinaweibopy" to install this sdk package to python, and also can download the sdk package, and put the webo.py with the py files I using to collect data into the same fiddler to use this sdk. I using the second method because I have issue pop this sdk. User can get the direction of how to use sdk via the weibo sdk wiki page [10], they provide many tutorial about how to use sdk on different environment not only for pyhton. For using Offical sdk, we need to use the "app\_key" and "app\_secret", we can find those code from the application page which I create the app apply before using my account. Those two codes are represent the user identity of who using the API, so weibo will ban the user's weibo account if they do something bad via weibo API because those two codes are directly link to user's weibo account. For getting the autorized for using the weibo API, I using Thinkgamer\_gyt's idea to get the authorized code [9], Weibo using OAuth 2 to check the user identity for using API. After the authorized page pump up, enter code which from the page url link which look like <https://api.weibo.com/oauth2/default.html?code=2024222384d5dc88316d21675259d73a>, and the code we need to enter is the string that after "code=" at the url link. ; then weibo will return an the access token for the API, then we using "Client" to activate the API, so we can get our target information via "Client".

After finishing open the API, the next step is to allocate the target feed page. For target the specific feed, we need to find out the id for each feed. Weibo is pretty tricky, they hide the real id and replace it as some codes at the feed URL, so we need to decode the Url to get the real feed id. [https://weibo.com/5305999252/Fy0sio7nQ?from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment](https://weibo.com/5305999252/Fy0sio7nQ?from=page_1005055305999252_profile&wvr=6&mod=weibotime&type=comment) this link is my target feed link, take a look on the link, we can find out the the code before the question mark is pretty much look like the encryption feed id. Xuebuyuan find out the encryption rule for

feed id [16], based on his idea, each four characters from back to front is a group as sixty binary, and switch those sixty binary to ten binary and then link them together. I using his code to decode the feed id at the ipython notebook, so user want to change their focus feed, they can enter the different codes from their focus feed's url, then they can get their feed id.

Once we get the feed's id, we can use this id to allocate the target feed at API. The next step is to get the target information we want to analysis from the feed. We are looking for the comments information from this page, so we need to know the code about the API port for our target information. Weibo provide a API port instructions to guide the user how to access different information via API [12], the access port we are looking for is comment. The code of comment port can not directly use at python, then use dot instead of slash for the comment for fit the python coding rule. Then following the access port guide, setting the parameter like feed's id, the number of page we want to have for the comment, and the number of comment we want to have for each comment page. For here, I only set up return 200 comments from page one because set limited usage of API for each account, so each account only get 2000 comments via API if the account is using free API connection, I don't want to use all the attempts chances at once. Then we use the access token which we got in the previous section to open the API and active the data port we set up for the target information. Finally set a variable to storage the data which return from the comment port.

The last step for collecting data is to storage the target data as txt file. All the data that related to comment are saved into the variable now, so if we want to get our target return object from this data set, we need to use the special code for different kind of category inside this data set; Weibo also provide a specific instruction for the code of return object [11]. The data we need for wrong information filtration are the content of comments and the user information for each comments, the numbers of follower and the number of friends for the user who post the comments, also we need to get the number of feed that post from the user who comments the feed. Then use special code "follower\_count", "FRIENDS\_count", and "statuses\_count" to get those information, and save them into the txt file for next step data vualization. The reason why I save my target date into txt file because of the coding knowledge shortage, I don't know how to using the data analysis model at 2.7 python version, so I decide to using Python 2.7 to collecting data and using python 3.5 to do the data analysis.

## 5.2 Data visualization

The data visualization in this paper will be simple and straight forward, because of even I have some idea to analysis the data, but I can not represent it due to coding knowledge shortage. The models I decide to use for this data visualization are matplotlib, nltk, wordcloud, pandas, numpy, jieba and codecs. First, open the content.txt file at python and using readlines and appends to create a list that content all the comments. I intend to use worldcloud to visualize the words that has most frequency on the comment list, and I find out the worldcloud don't support chineeses really well, so I use the module "jieba" to reproduce the comments list. Jieba is the best module to support Word Segmentation, wen can using

this module to pick up the words which most frequently appear at the comment list[4]. I using FontTian's formatting as my main structure of jieba code [4], also we can add some new word into our word list and use it to make the jieba module can be able to indentify the new world, and we use "stopwords\_path" to filter the common word like "hello", and the we are using outside txt source which is Chinese vocabulary words out file as our stop word dictionary [1]. And during using this stop word file inside the jieba code, we have to encoding the file to "utf8" formatting, otherwise it will have some error that the jieba module can not distinguish the content inside the stop word file; also we need to set up the right font for the wordcloud by using the ttc file from the fonts document on the computer, if the font use on wordcloud doesn't support chinese, the final result will be a retangle for each word instand of actually Chinese.

The first outcome I got from the word cloud is look like this:

[Figure 2 about here.]

So now people can really quick have a pretty idea that what is the main trend of the comments, and what is all the comment talking about. But we can still find some noise inside the plt, like "ffh" which mean replyTh it doesn't contain any meaning, so we can add this word to our stop words list, then we try to create world cloud again, the new plot is look like this:

[Figure 3 about here.]

It seems more meaniful than before, and contain more information than before. So we can see that their lot of words like "ffh" inside the comments data is useless, then we can add those word into our stop words list to rip it up. Also we can use this method to filter the spammer's information, so the user can see the true information from the world cloud.

Then for detecting the spammer, analysis the user who post the comments is very important. To visualize the user data, using readlines to open the each txt file, and using solit and strip to reproduce the data formatting. Then putting all the data into the dataframe via pandas modules. When I look at the data type in dataframe, I found out that the data type inside the dataframe for each catory is not number, then I use astype to change the data type to number for calculator. Here is the three histograms I plot out for the the Feeds, Friends, and followers data:

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

The result is pretty interesting and surprise for me. The data represent those three histograms are pretty obviously, even the number of my sample is only 200 comments because of the limitation of weibo API. We can see that all the histograms are right-skewed distribution, according to the definition of histogram, the mean at right skewed distribution is the peak of the right side [1]. It doesn't look like normal because of the curve is not bell shape, if this data is from the real commets which mean post by the actual user not spammer, the curve of this three histograms will looks like normal distribution. The peal of those three histograms show us the majority of those three elements, for the number of friends, the majority is between 0 to 250; for the majority of feeds is between 0 to 1800, and the majority for followers is between 0 to 200. we

can see most majority are fall into the first interval, which mean it represent the friends, followers, and feed from the accounts I pick are pretty much same type of fake account. So it is pretty luck that I can dig out so much spammer account with small number of simple, so there no doubt that their are lot of spammers involve into this feed's comment because the comments post by the majority accounts are pretty much same type of account.

## 6 FUTURE WORK

The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more. Furthermore, using machine learning to train a model that can recognize the most frequency word that spammer use to post the wrong information, and use it the find out the spammer on the comment list. Finally, according to the comment list data analysis, save the user name which are define as spammer in database, and then remove the comments that related to those user in the comments content list, use this new comments content list to create wordcloud to represent the true trend and focus point for user's target feed. Moreover, the spammer data on the database will be cumulative, so the application analysis more fee, the spammer list will be increase, then each time the application can remove the comments which post by the spammer on the spammer list before analysis the spammer on the comments list which can improve the precision ratio of eliminate spammer information a lot. I think big data is based on the data precipitation, so most big data application won't have good performance at the begin because lack of data, when the application produce and save data until certain level, the performance the application will be increase.

## 7 CONCLUSIONS

The power of social opinion not only effect the ciber world, but also have great impact on real world. Therefore, it is really important to let the user in ciber world getting right information for the content that they are interesting in; the best way to achieve this gold is using application that base on big data analysis to filtering the wrong information that post by the spammer. There lot of ways to filtering the wrong information, but the collecting related data are always same, because no matter using which way to analysis the data, getting data is top priority than any things. I believe current technology can support big data storage really well, when the data storage reach certain amount, we can use it to decontaminate the ciber world, and maintain the ciber world envirnment that allow

people gain real information and create real social relationship on it. Therefore, improve the accuracy and adaptation for spammer will be meaningful to investigated.

## 8 ACKNOWLEDGEMENT

I would like to take this chance to thanks to my tutor Miao, in process on reviewing my paper, he gave me many useful comments and advises. Finally, I would love to thanks my friends who working at IT company give me many idea about how spammer work.

## REFERENCES

- [1] ASQ. 2017. Typical Histogram Shapes and What They Mean. (2017). <http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html> [Online; accessed 1-Dec-2017].
- [2] Christina Hills. 2017. DIFFERENCE BETWEEN SPAMMERS AND HACKERS. (2017). <https://websitecreationworkshop.com/blog/wordpress-tips/difference-spammers-hackers/> [Online; accessed 1-Dec-2017].
- [3] Eddie Lee. 2016. Next Magazine to pay BaWang shampoo makers HK\$3 million compensation for defamation. (2016). <http://www.scmp.com/news/hong-kong/law-crime/article/1951576/next-magazine-pay-bawang-shampoo-makers-hk3-million> [Online; accessed 1-Dec-2017].
- [4] fxsjy. 2017. jieba. (2017). <https://github.com/fxsjy/jieba> [Online; accessed 1-Dec-2017].
- [5] Garrett Brown and Travis Howe and Micheal Ihbe and Atul Prakash and Kevin Borders. 2017. Social Networks and Context-Aware Spam. (2017). [http://web.eecs.umich.edu/~aprakash/papers/cscw08\\_socialnetworkspam.pdf](http://web.eecs.umich.edu/~aprakash/papers/cscw08_socialnetworkspam.pdf) [Online; accessed 1-Dec-2017].
- [6] isocialfame. 2017. Buy Real Facebook Page Likes+Followers (Business Pages). (2017). <https://isocialfame.com/collections/facebook-marketing/products/buy-facebook-fan-page-likes?variant=509391732763> [Online; accessed 1-Dec-2017].
- [7] KR Canini and B Suh and PL Pirolli. 2017. Finding Credible Information Sources in Social Networks Based on Content and Social Structure. (2017). <http://www.parc.com/content/attachments/finding-credible-information-preprint.pdf> [Online; accessed 1-Dec-2017].
- [8] michaelliao. 2017. sdk. (2017). <http://github.liaoxuefeng.com/sinaweibopy/> [Online; accessed 1-Dec-2017].
- [9] Thinkgamer. 2017. Weibo API using guide. (2017). <http://blog.csdn.net/gamer-gyt/article/details/51839159> [Online; accessed 1-Dec-2017].
- [10] Weibo. 2017. SDK. (2017). [http://open.weibo.com/wiki/SDK#Python\\_SDK](http://open.weibo.com/wiki/SDK#Python_SDK) [Online; accessed 1-Dec-2017].
- [11] weibo. 2017. weibo API return object code. (2017). <http://open.weibo.com/wiki/> [Online; accessed 1-Dec-2017].
- [12] weibo. 2017. weibo API wiki. (2017). <http://open.weibo.com/wiki/weiboAPI> [Online; accessed 1-Dec-2017].
- [13] weibo. 2017. Weibo application. (2017). <http://open.weibo.com/apps> [Online; accessed 1-Dec-2017].
- [14] Xu Chang. 2013. Detecting collusive spammers in online review communities. (2013). <http://www.ixueshu.com/document/43b579eeddbe46b2318947a18e7f9386.html> [Online; accessed 1-Dec-2017].
- [15] xudong luo and pin liu. 2011. analysis the spread of spammer. (2011). <http://www.ixueshu.com/document/43b579eeddbe46b2318947a18e7f9386.html> [Online; accessed 1-Dec-2017].
- [16] xuebuyuan. 2007. get mid. (2007). <http://www.xuebuyuan.com/1874313.html> [Online; accessed 1-Dec-2017].
- [17] zhibin liu and lanhua deng. 2017. analysis the credible in network information. (2017). <http://www.ixueshu.com/document/1453923d337e1742318947a18e7f9386.html> [Online; accessed 1-Dec-2017].

LIST OF FIGURES



联合国 V

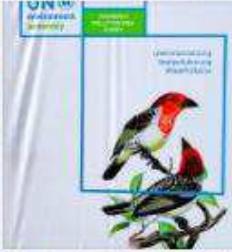
12月3日 22:37 来自 iPhone 8 Plus



和@微公益一起近距离关注 2017 #联合国环境大会# @联合国环境规划署 @李晨

@微公益 V

#联合国环境大会# 即将开幕！微公益带你一起走进肯尼亚，迈向#零污染地球#！@联合国环境规划署 @熊猫守护者 @微环保



12月3日 17:16 来自 荣耀9 美得有声有色

126 | 15 | 76

☆ 收藏

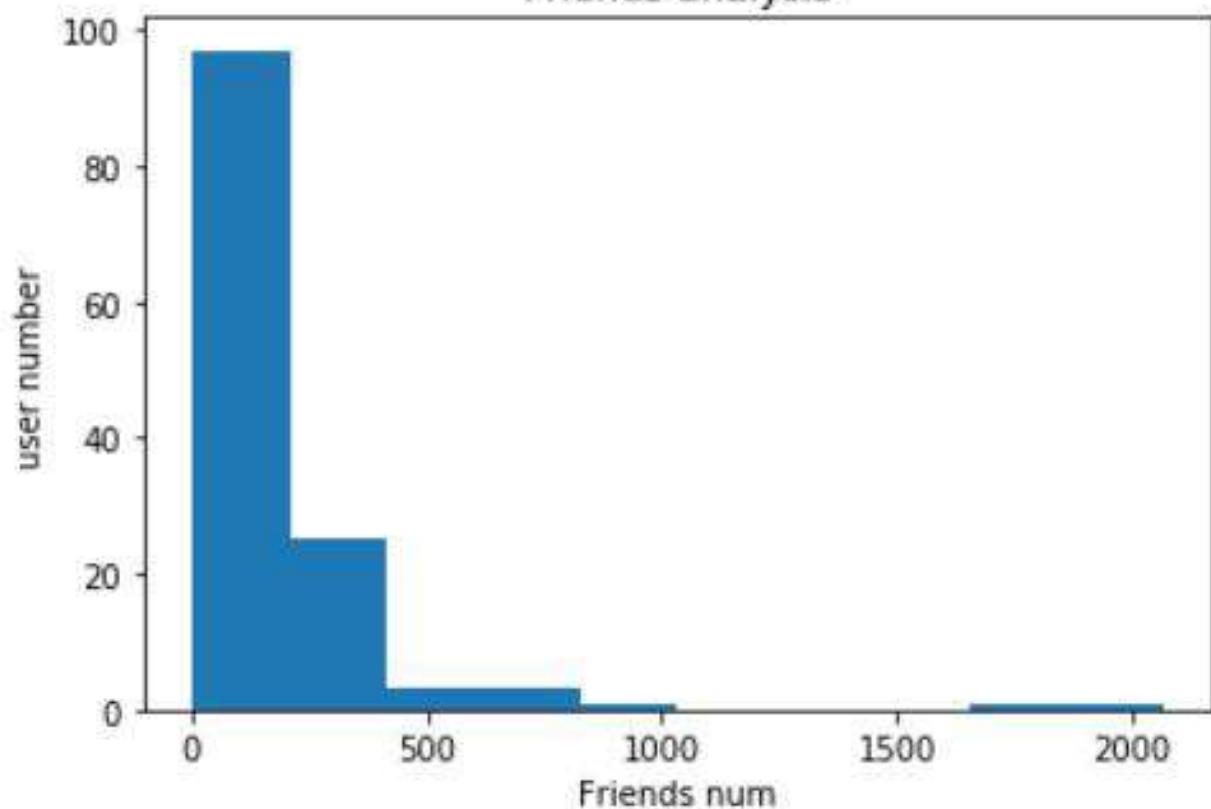
74

48

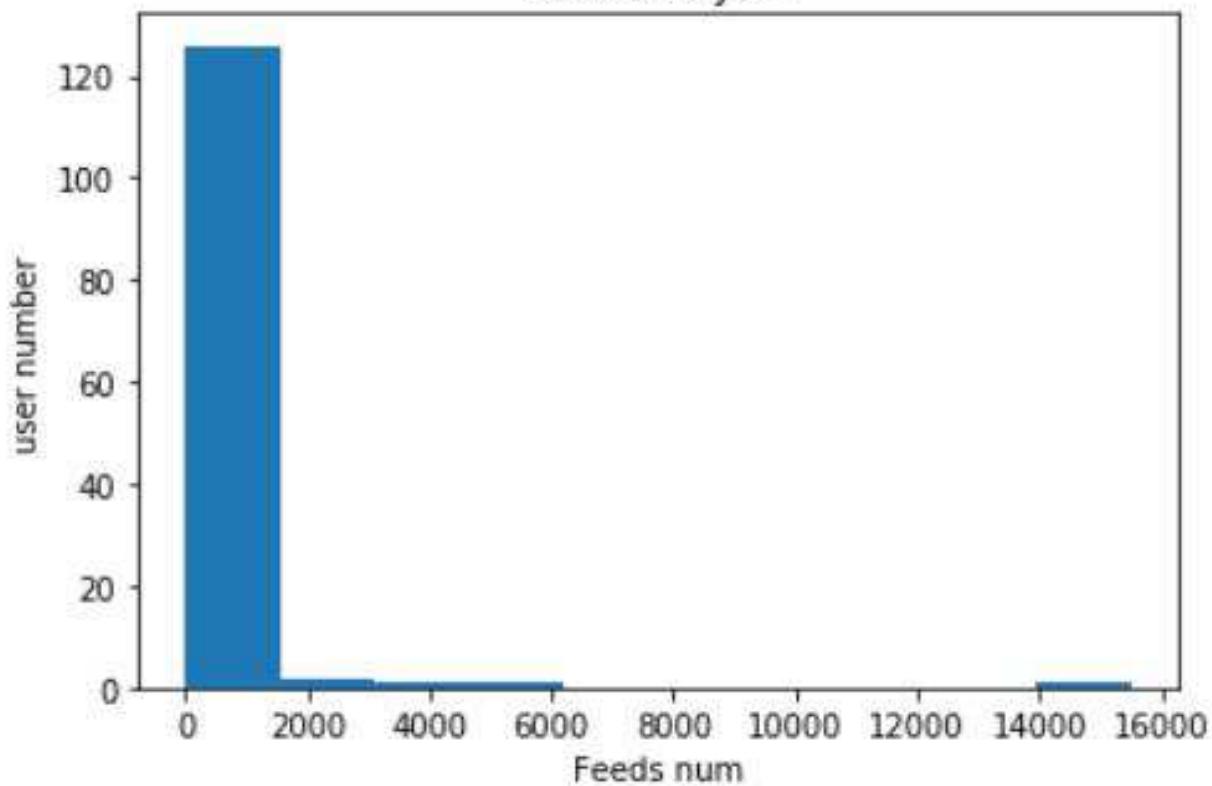
349



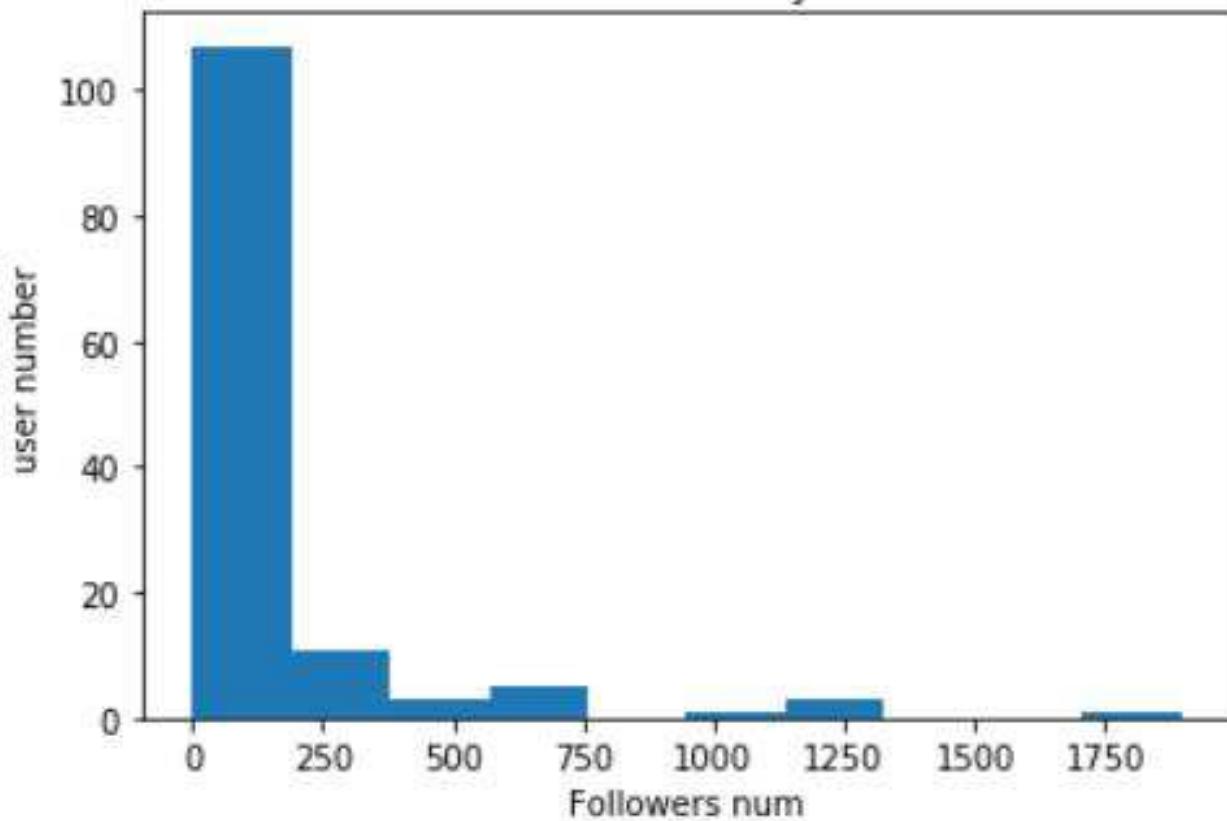
Friends analysis



Feed analysis



### Followers analysis



```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "intro01"
Warning--I didn't find a database entry for "intro02"
Warning--I didn't find a database entry for "target01"
(There were 3 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.31.55] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Citation 'intro01' on page 1 undefined on input line
Missing character: ""
Citation 'intro02' on page 1 undefined on input line
Citation 'target01' on page 2 undefined on input lin
Missing character: ""
```

```
Missing character: ""
There were undefined citations.
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
'h' float specifier changed to 'ht'.
Typesetting of "report.tex" completed in 1.0s.
./README.yml
8:81      error    line too long (84 > 80 characters) (line-length)
9:81      error    line too long (85 > 80 characters) (line-length)
10:81     error    line too long (85 > 80 characters) (line-length)
11:81     error    line too long (82 > 80 characters) (line-length)
22:1      error    trailing spaces (trailing-spaces)
25:15     error    too many spaces after colon (colons)
25:81     error    line too long (691 > 80 characters) (line-length)
41:81     error    line too long (96 > 80 characters) (line-length)
42:81     error    line too long (688 > 80 characters) (line-length)
```

---

## Compliance Report

---

```
name: Ni, Juan
hid: 107
paper1: Oct 22 1800 100%
paper2: 100%
project: Dec 08 0600 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
11
wc 107 project 11 4612 report.tex
```

wc 107 project 11 4697 report.pdf  
wc 107 project 11 2025 report.bib

find "

---

37: The social opinion trend at social networking community will drive personal and even company decision, then some trends might harm someone's benefit because the power of social opinion trend is so powerful. Then people hired spammers to spread wrong information that lead the trend become advantage for them, but the users are become victim because they will make wrong decision because of the trend is control by someone on purpose. "Brown showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks." \cite{adv:01}, every spammer post must for some reason that beneficial for their employer, the most famous case for spammers the shampoo case at 2010. "BaWang shampoo" is the most famous shampoo at China which advertised by super star Jackie Chan, "Next Magazine" post a fake news claimed that using "BaWang shampoo" could cause cancer \cite{bawang:01}. I clear remember at that time, almost all social websites post new claim "BaWang shampoo" is harmful at the same time without any authority judgment, and they put this new at the headline position to abstract user's eye-ball. Even the authority department proof this new is unreliable, the business reputation of "BaWang shampoo" had been damaged, lot of people around me stop using this shampoo any more. This case seems have no spammers involved, but actually the spammers for this case is social networking websites themselves instead of single person. The reason why they post slander is because they can get benefit from other shampoo companies in china, other shampoo companies can have more sales because the market-share of "BaWang shampoo" will be decrease at this case.

39: The other reason why spammers getting so popular at social networking website is the operating cost of spammers is supper low. Try searching "buy Facebook like" at Google, and there are over hundred million results come up. And the price of buying like and followers from that website is pretty low, so spammers can get an account with 1000 followers which look like a real account for only 5 dollar \cite{buy:01}. So spammer can create thousands this kind of fake account for posting the wrong information at social networking website, and the detected system is hard to find out those kind of account is real account or Zombie account those accounts have actually followers and like, even feeds. Then spammers can use script to post batch of wrong information via

those account. Furthermore, the cost for post batch of wrong information is unbelievable low; according to the internal information which provided from my friend who working at a IT company, there two ways to post wrong information at social networking website, first one is money reward system, second one is posting AI. For the reward system, a professional spammer company usually have about 10 teams, each team has 500 people; They use reward instead of constant salary, each feed related to the order topic is worth 0.5 Chinese dollar(equal to 5 cent in dollar), and each comment on the target feed is worth 0.2 Chinese dollar( equal to 3 cent in dollar), and the price of long text post is negotiated. The spammers accounts are provide from the company, the price for those account is also low; Most social networking websites only require email address for registered, then they buy email accounts from the retail like 100 Chinese dollar(equal to 15 dollar) for ten thousand accounts, and using script to register social websites accounts and making follow with each fake accounts. This is how they operate the spammer, now most social networking websites register require phone number verify, then they working with local sim card retail which infinity phone number on hand, but the price for each fake account is increase a lot like 3 Chinese dollar(like 50cent) per one, but still pretty cheap compare with other advertise way. Cheap labor force and the development of script technology rising the spammer company, but the lower the user experience at social websites because lot of trash information full of the social websites, people are hard to see the true at social websites any more. According to the network sites worldwide ranking[7], we can see WeChat has almost twice active user than Weibo, and WeChat is the most popular social networking app in China because spammer can not do any at that app. In WeChat, they post feed at the module which call "Friends circle", the feed formatting is pretty similar as weibo, but the app is semi-closed which mean it is complete private, user can only see their friend's post and the comment, no retweet allowed, and if someone who not in the user friend's list comment at user's feeds, user can not see that guys comments. Plenty of users quit traditional social networking website, and the semi-closed social networking app is getting popular, so the adverse eff of spammer is not only for the spam target but also for the platform. If we let spammer keep development, and don't have any way to filter their information, the traditional social websites will die soon.

- 43: According to the worldwide statistics data, "Sina Weibo" has 368 millions active users which more than 328 millions of twitter active user \cite{target01} , so I would like to using "Sina Weibo" as my investigate target instead of using Twitter. The

other reason why I choose "Sina Weibo" as my investigate target is because I'm familiar with Chinese culture and I have been using "Sina Weibo" for more than 8 years. I think my knowledge about "Sina Weibo" will help me a lot at this project and better understand how spammers works at "weibo". The page frame at "weibo" is pretty similar to Twitter [figure1].

- 50: The four buttons under each feed are "collect, retweet, reply, like", and the capability for each button is same as twitter. When user click into the "rely" button, user can see all the comments related to the current feed, and sort them by the amount of "like" that comments get from other user. The only things differen at "weibo" is user not only can see the comments but also can see the retweet information, twitter only allow user to see who retweet the feed. Then user can see the retweet's comments and sort the list by the retweet's times of the retweet feed. So people would love to check the retweet list to see which famous person retweet the feed, and what comment they put into the retweet. Spammer control the public opinion trends by putting wrong information that doesn't represent real public opinion into the comment for some hot feed, they utilize user's habit to reach their goal.
- 60: from=page\\_1005055305999252\\_profile&wvr=6&mod=weibotime&type=comment} this feed to investigated the comment content. The person send this feed is my favorite gaming live streaming player, his name is LuBen Wei. He is the most popular gaming live streaming in China, there are over four million audiences what his playing game every night. Moreover, this is his second account, so he always post some feeds that can not be post at his official account at Weibo, but there are still thirty thousand comments under this feed , the number of comments at this feed even more than the comments under every signaler feed from Donald J. Trump's account. And the feed's content is he complain about the cheating case, he announced that he never cheating at "PlayerUnknown's Battlegrounds", he claim that the rumor about his cheating is come from the spammers. After he post this feed, this feed became the top 1 hot feed at the feed ranking at Webo, and most comments under this feed are abuse him cheating. So I though there are must be spammer working under this feed, the comments at a feed from a gaming live streaming player's second account is more than the comments from United states's president's feed which is so ridiculous. Therefor I think there must be spammer involved into this feed, that is how we pick up the feed which involved spammer in social networking website. If a feed has unusual comments and likes amount compare with other feed post from the owner, that feed have huge possibility that involved spammer work.

- 62: First of all, Weibo require we use Weibo API with authentication, so we need to create a personal application first at the weibo application apply page \cite{method:01}. Then the weibo official suggest us to use SDK to access the the API, so I came to the sdk websites \cite{method:03} to get the Weibo SDK package. Normally can just type "pip install sinaweibopy" to install this sdk package to python, and also can download the sdk package, and put the webo.py with the py files I using to collect data into the same fidder to use this sdk. I using the second method becuase I have issue pop this sdk. User can get the dirction of how to use sdk via the weibo sdk wiki page \cite{method:04}, they provide many tutorial about how to use sdk on different environment not only for pyhton. For using Offical sdk, we need to use the "app\\_key" and "app\\_secret", we can find those code from the application page which I create the app apply before using my account. Those two codes are represent the user identity of who using the API, so weibo will ban the user's weibo account if they do something bad via weibo API because those two codes are directly link to user's weibo account. For getting the autorized for using the weibo API, I using Thinkgamer\\_ggt's idea to get the authorized code \cite{method:05}, Weibo using OAuth 2 to check the user identity for using API. After the authorized page pump up, enter code which from the page url link which look like
- 63: \url{https://api.weibo.com/oauth2/default.html?code=2024222384d5dc88316d21675259d73a}, and the code we need to enter is the string that after "code=" at the url link.
- 64: ; then weibo will return an the access token for the API, then we using "Client" to activate the API, so we can get our target information via "Client".
- 72: The last step for collecting data is to storage the target data as txt file. All the data that related to comment are saved into the variable now, so if we want to get our target return object from this data set, we need to use the special code for different kind of category inside this data set; Weibo also provide a specific instruction for the code of return object \cite{method:07}. The data we need for wrong information filtration are the content of comments and the user information for each comments, the numbers of follower and the number of friends for the user who post the comments, also we need to get the number of feed that post from the user who comments the feed. Then use special code "follower\\_count", "FRIENDS\\_count", and "statuses\\_count" to get those information, and save them into the txt file for next step

data viualization. The reason why I save my target date into txt file because of the coding knowledge shortage, I don't know how to using the data analysis model at 2.7 python version, so I decide to using Python 2.7 to collecting data and using python 3.5 to do the data analysis.

75: The data visualization in this paper will be simple and straight forward, because of even I have some idea to analysis the data, but I can not represent it due to coding knowledge shortage. The models I decide to use for this data visualization are matplotlib, nltk, wordcloud, pandas, numpy, jieba and codecs. First, open the content.txt file at python and using readlines and appends to create a list that content all the comments. I intend to use wordcloud to visualize the words that has most frequency on the comment list, and I find out the wordcloud don't support chineses really well, so I use the module "jieba" to reproduce the comments list. Jieba is the best module to support Word Segmentation, wen can using this module to pick up the words which most frequently appear at the comment list\cite{method:08}. I using FontTian's formatting as my main structure of jieba code \cite{method:08}, also we can add some new word into our word list and use it to make the jieb module can be able to indentify the new world, and we use "stopwords\\_path" to filter the common word like "hello", and the we are using outside txt source which is Chinese vocabulary words out file as our stop word dictionary \cite{method:10}. And during using this stop word file inside the jieba code, we have to encoding the file to "utf8" formatting, otherwise it will have some error that the jieba module can not distinguish the content inside the stop word file; also we need to set up the right font for the wordcloud by using the ttc file from the fonts document on the computer, if the font use on wordcloud doesn't support chinese, the final result will be a retangle for each word instand of actually Chinese.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

5: \input{format/i523}

```
passed: True
```

```
find input{format/final}
```

---

```
passed: False
```

```
floats
```

---

```
44: \begin{figure}[h]
45: \includegraphics[width=\columnwidth]{1.jpg}
78: \begin{figure}[h]
79: \includegraphics[width=\columnwidth]{2.jpg}
82: \begin{figure}[h]
83: \includegraphics[width=\columnwidth]{3.jpg}
89: \begin{figure}[h]
90: \includegraphics[width=\columnwidth]{4.jpg}
92: \begin{figure}[h]
93: \includegraphics[width=\columnwidth]{5.jpg}
95: \begin{figure}[h]
96: \includegraphics[width=\columnwidth]{6.jpg}
```

```
figures 6
```

```
tables 0
```

```
includegraphics 6
```

```
labels 0
```

```
refs 0
```

```
floats 6
```

```
True : ref check passed: (refs >= figures + tables)
```

```
True : label check passed: (refs >= figures + tables)
```

```
True : include graphics passed: (figures >= includegraphics)
```

```
True : check if all figures are refered to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

```
[width=1.0\columnwidth]
```

```
do not change the number to a smaller fraction
```

```
find textwidth
```

---

passed: True

below\_check

---

WARNING: table and above may be used improperly

104: The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more.

WARNING: code and above may be used improperly

104: The above sections bring out the idea about how to detect the spammer, and the method is pretty sample because of it just to use for proof my idea is feasible. To rise the filter wrong information to the big data level, we can using database to storage our data instead of storage the data into a txt files. We can make a connection between API and mysql, and the programming will automatically storage the data inside each table for different data category. Also the authorized code can be automatically get from the authorized page, so user no need to enter it by hand. All the analysis will be integration into one application, so user only need to copy and paste the feed link that they want to see the true for that feed, and the application will decode the feed id and collecting all kind of data into the mysql database. For collecting huge size of data like over ten thousand comments, we can using different virtual machine to get data from the API, so we don't need to worry about the daily API usage any more.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "intro01"
Warning--I didn't find a database entry for "intro02"
Warning--I didn't find a database entry for "target01"
(There were 3 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

```
382: % editor =      "",  
383: % number =     "",  
385: % type =       "",  
386: % chapter =    "",  
387: % pages =      "",  
388: % address =    "",  
389: % month =      "",  
  
passed: False
```

ascii

---

```
non ascii found 65292
non ascii found 65306
non ascii found 8220
non ascii found 8221
non ascii found 65306
non ascii found 65306
non ascii found 22238
non ascii found 22797
non ascii found 65292
non ascii found 22238
non ascii found 22797
```

```
=====
The following tests are optional
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
-----
```

```
passed: True
cites should have a space before \cite{} but not before the {
```

```
find cite {
-----
```

```
passed: True
```

# The Importance of Data Sharing and Replication, But What About Data Archiving?

J. Robert Langlois

Indiana University

Bloomington, IN 47408, USA

langloir@umail.iu.edu

## ABSTRACT

With the increase of digital information, scientists have faced many challenges when it comes to the topic of big data management, including data archiving and data sharing. While it is unproblematic to share and archive quantitative data, qualitative data remains a puzzle that social scientists need to solve when it comes to what data to share, where to house the data, who will pay to store the data, how long the data should be kept for, etc. Many researchers are skeptical to engage in the practice of sharing digital information due to privacy concern, fear of stigmatization, the problem of funding, repository of data, transparency, and so forth. While it is important to keep these challenges in mind, it is critical to take a look at the different advantages of data sharing and data archiving.

## KEYWORDS

i523, HID325, Data Sharing and Data Archiving

## 1 INTRODUCTION

Nowadays, many fields are witnessing a large influx of data due to the increase usage of technology. The digital data generated from scientific research is integral to the advancement of different scientific fields. While these data sets exist in abundant quantities, one challenge that many fields face is the lack of and/or prohibition of data being shared among researchers. Data sharing and its subsequent replication is a subject matter that is in dispute within in the sciences [13]. This is a significant area of contention in the United States; however, in other countries like the United Kingdom (U.K.), this issue has been addressed by making data sharing a matter of great importance; so much so, that the Joint Information Systems Committee of the U.K. made "data-sharing a priority, and has helped to establish a Digital Curation Centrefit to be national focus for research and development into data issues" [16]. On one hand, opponents of data sharing are skeptical about this practice due to privacy concerns, fears of stigmatization, funding problems, repositories for data, transparency, etc. On the other hand, some researchers are open to data sharing because it allows their work to be reviewed and creates opportunities to further their findings; however, the actual practice is stunted by researchers concerns [15]. Proponents of data sharing continue to advocate for an open access policy that would allow data to quickly respond to societal problems and crises, as well as advance the sciences. While it is as important to keep in mind the different challenges to data sharing, like data archiving, sharing data among fellow scientists can be very beneficial, not only can this practice help to maximize profits, make new discoveries, and respond to crises more quickly, but also it can play a vital role in advancing science and research.

## 2 THE RELEVANCE OF DATA SHARING AND DATA REPLICATION

### 2.1 Define Big Data, Data Sharing, and replication

Big data "is data that exceeds the processing capacity of traditional databases. The data is too big to be processed by a single machine. New and innovative methods are required to process and store such large volumes of data" [10]. The data sets are so voluminous and complex that traditional way of data processing application software are insufficient to deal with them.

Data sharing can be understood as the ability to share research findings with multiple users. This technique implies that the digital information is being archived in one or multiple servers in the network and that there is other technique to prevent this information from being altered by two or more users at the same time. It is the practice of making data used for scholarly research available to other investigators. Data sharing is nothing but making the data available for other users to use for the common good of society. [1].

Data replication is the process of copying data from one location to another. The technology helps an organization to possess up-to-date copies of its data in the event of a disaster."Replication can take place over a storage area network, local area network or local wide area network, as well as to the cloud. For disaster recovery (DR) purposes, replication typically occurs between a primary storage location and a secondary offsite location" [2].

### 2.2 The Advantages of Big Data

We cannot talk about the advantages of Big data without quickly acknowledge that its numerous challenges, which include data collection, data storage, data analysis, search, sharing, transfer, visualization, etc.

Big data analysis presents numerous advantages. For instance, it helps businesses to increase their productivity. This has done through a process of analyzing raw data that produces information that identifies trends and patterns that will help businesses make cost effective decisions. It is also helpful in aiding government agencies to improve public sector administration, and assists global organizations in analyzing information that has wide-reaching impact on the world. The information produced by big data can help medical professionals to detect diseases in earlier stages. Some other advantages of big data analysis is present in many different areas, such as: smart grids, which monitor and control electricity use; traffic management systems, which provide information about transportation infrastructure like roads and highways, mass transit, construction, and traffic congestion; retail by studying customer

purchasing behavior to improve store layout and marketing; payment processing by helping to detect fraudulent activity, etc [23]. Data is being collected everywhere due to the use of the internet; therefore, businesses and scientists are trying to make the maximum profits from it. Data sharing, for instance, can help scientist to respond to epidemic and crisis in a quickly manner.

**2.2.1 Data Sharing Helps to Respond to Crisis Quicker.** Sharing data among fellow scientists and researchers is crucial. This process can help to respond to crisis quicker. By sharing (digital) data, researchers do not always have to start from scratch when they are responding to societal problems, such as medical epidemics, economic instability issues, natural disasters, etc. As previously mentioned, an important aspect of data sharing is its ability to be used to respond to and help expeditiously resolve societal issues. As found in [26], data sharing is encouraged among fellow researchers and scientists to quickly respond to outbreaks. The authors centered their arguments around the rise of Ebola back in April 2015 that raised serious panic around the whole world due to the dangerous impacts of the disease that could result in death from just one exposure. They explained how the rapid availability of research data had facilitated a more amenable response time to the rapidly spreading threat of Ebola. The data accessed allowed it to be determined that the virus had circulated from Guinea to Sierra Leone and that it was being sustained by person-to-person contact.

The fact that the data was recoverable from the GenBank, a public database, allowed researchers ready access and assisted with tracking the source of the deadly virus; thus, leading to the advocacy for the sharing of data among researchers to allow for quicker responses to life-threatening crises. This is one example of how data sharing can have a crucial impact on the response time researchers have when responding to disasters that have a global impact.

Moreover, [25] wrote about the problem that public health policies faced when it came to responding to outbreaks like Ebola. He explained how bureaucracy and a lack of record keeping often delayed the ability of scientists to respond. A lack of collaboration among researchers can hinder progress when scientists need to respond to crises. Not only does data sharing help scientists to respond to and help provide critical solutions to outbreaks like Ebola, but access to the data can also contribute to the advancement of science through data replication. Thus, the importance to encourage this practice among fellow researchers.

Certain research studies have supported the idea that big data allows for real time tracking of diseases and the development, prediction of outbreaks, and facilitates the development of personalized healthcare. Big data can also be used to maximize profits in many disciplines, including healthcare if harnessed properly [24]. "By harnessing big data, businesses gain many advantages, including increased operational efficiency, informed strategic direction, improved customer service, new products, and new customers and markets" [11]. While data exists in huge quantities in many fields, including the health care field, individual privacy concerns remain a big problem that policymakers have to tackle to meet current trends in data collection. Improved methods of protecting very personal, private and sensitive health information is needed in order to allow

for safe, necessary and adequate access to protected health information within the health care industry. Without proper policies related to data use, access, and protection, this big data potential can not be realized [17]. Data sharing and replication contribute to increase the availability of the amounts of digital information and make the access to information easier.

**2.2.2 Data Sharing and Replication Contribute to Increase the Availability of Digital Information.** Furthermore, data sharing and replication (the availability of multiple copies of a data set to different users) is an important practice that plays a crucial role in the advancing of the sciences. The way sciences grow is through scaffolding, which means that one must rely on the work of previously published research to come up with new findings; thus, further supporting the necessity for a collaborative effort among researchers and scientists. Data sharing has been shown to help spot errors in research. In 2013, for example, a graduate student pointed out a calculation error made by two Harvard professors. This discovery was only possible because the professors shared a spreadsheet of their research findings with the particular student [13]. In this case, and possibly in many other cases, data sharing has helped to build community among fellow researchers, uncover honest mistakes and, in worst-case scenarios, expose possible fraud. Another researcher made a series of observations about the relevance of sharing data and asked many poignant questions regarding content, parameters and the necessity for sharing information. One observation she made was that "science progressed for centuries without data sharing policies and then questioned, why is data sharing deemed so important to scientific progress now?" [6]. She challenged the notion of free and unhindered distribution of data and cautioned that preliminary questions must first be answered to determine what data to share, how much and in what context data should be released to advance the changes in sciences. Her point was that the data should not stand alone, but rather it ought to be accurately defined and contextualized within the means that identified, developed and synthesized the data into usable information. While Borgman's scrutiny has its place in the argument regarding the absence of policies! which she argues ought to be based on accurately defining the data parameters! and their ability to facilitate data sharing, it is also important to acknowledge that the various scientific fields have been faced with different types of data and challenges.

Nowadays, scientists are being bombarded with an abundant presence of digital data, which has made it difficult to manage and store, and far more difficult to compare data exchanges to what it was centuries ago. If digital data that is generated through research is not being replicated, the world of science will face far more challenges in the years ahead due to a lack of evolving data to build upon. Data replication involves open access to data so that researchers can continually study, analyze, and make new discoveries about existing data. Now, if data sharing opens the door to the replication of scientific research and advancement, then why is there so much opposition to such a practice? Without replication, the sciences may become stagnant in their advancement of theories and potential solutions to global problems. Not only data sharing makes access to information easier, but also helps to mitigate/eliminate unnecessary cost.

**2.2.3 Data Sharing Help Reduce Unnecessary Cost**. Another reason to encourage data sharing is that data sharing help to reduce unnecessary cost. As digital information is made available in the cloud system, researchers will be able to access the same type of information at different locations. "Data sharing is driven by the need to maintain more accurate and up-to-date spatial databases, but at the same time reduce data acquisition and maintenance costs" [21]. If data becomes more accessible, not only this will contribute to lowering the cost to access data, but also will it encourage researchers in their endeavors to respond to social outbreak quicker. Data sharing can also play a crucial role in advancing science, which is our next point.

**2.2.4 Data Sharing Advances the Science.** Contrarily to what many might think, the advancement science occurs through replication of existing findings; scientists must rely on the work of previous researchers to make new discoveries. Just like human being cannot live in vacuum; the way science develop is through collaborative effort among fellow scientists."Placing research data online allows instantaneous access by a globally dispersed group of researchers to share, understand, and synthesize results. This aggregation and synthesis provide an opportunity for insight, progress, and that uniquely human quest for larger understanding. Data repositories also allow for the publication of previously hidden negative data, essentially experiments that didn't work" [3]. One advantage of this practice is that by sharing their work, scientists will be able to spout errors that previous researchers have made, reveal fraud, build community, etc [13]. As found in [5] policy makers must develop policies that explain how to embark in this process. If data are not being replicated, the world of science will face far more challenges in the future. Data replication involves open access to data so that researchers can continue to study, analyze, and make new conclusion about existing data. Now, if data sharing allows the replication of the sciences, why are many people opposed such practice?

### 3 BLOCKS TO DATA SHARING

There are numerous barriers to data sharing. One of the barriers to data sharing is transparency.

#### 3.1 Transparency Issue

Some researchers fear that their work is going to be poached and that they will not get credit for their findings, so they hold on to it and do not disclose it. The concerns of obscurity and/or credit being assigned to other researchers who might advance the original researcher's findings will cause a serious reluctance to sharing data. As found in [13], the term data parasites is used to describe the practice of utilizing data without giving proper credit to prior publishers. Thus, to overcome this challenge, there should be honesty and allowance for intellectual probity; a sort of fair play between researchers where appropriate credit would be given. In addition to giving appropriate credit for findings, another important aspect of disclosure is the appropriate compensation for the release of intellectual property. Oftentimes, researcher have sacrificed their time, income potential, energies, and relationships to do the work they do; incentive needs to be provided to encourage the sharing of their findings that they have worked hard to develop. "The call for

transparency is not new, of course. Rather the emphasis is on access to data in a usable format, which can work to create value to individuals" [23]. Access to digital information can engage individuals, invite scrutiny, and expose misuses of data.

Another concern pointed out by this author is that data sharing may open the door for data analysts to disprove and/or scrutinize the work of the data producers. A potential solution that he proposed to this issue was co-authorship. He believed this would discourage the misuse of data by allowing collaboration among researchers [13]. Another researcher also asserted that researchers ought to agree on the standards of practice needed to responsibly share data. She advocated that both data and its means of publication deserve equal status in scholarly communications to determine how to cite data in non-trivial ways [6]. If data sharing and collaboration among researchers is to be effective, there need to be norms and regulations of how to do so. Collaboration among scientists can be a good thing to help mitigate and even eradicate the sense of fear that many researchers have in sharing their findings and the methodologies used to produce them. This leads us to our next potential block and challenge to data sharing, privacy concern.

### 3.2 The Problem of Individuals' Privacy

Another barrier to data sharing, specifically in the healthcare field, involves the protection of patient privacy; a lack thereof can lead to stigmatization and potentially hamper patients participation in healthcare research and treatment. "Privacy is a major concern in outsourced data. recently some controversies have revealed how some security agencies are using data generated by individuals for their own benefits without permission. Therefore, policies that cover all user privacy concerns should be developed. Furthermore, rule violators should be identified and users data should not be misused or leaked" [11]. In so doing, individuals will feel more at ease to engage in the process of sharing information for the benefits of everyone.

As [26] highlighted, some uncertainties that are involved data sharing, like whether data belongs to public or private domains. Still, another barrier is patient consent and their ability to fully understand how their participation can make them vulnerable to being potentially shunned and ostracized in their community based on their diagnosis and/or treatment. The researchers advocated for the responsible sharing of pertinent information among researchers to avoid this problem. It should also be mentioned that preclusion to the sharing of unnecessary information would also weaken the barriers to data sharing. Although data sharing is important, particularly during a medical outbreak, researchers ought to do their best to protect patient privacy to avoid any threat of stigmatization or isolation of patients. Rigorous ethical standards should be applied to safeguard patients' privacy and dignity to allow for easier sharing of relevant data [26]. Shelton (2011) advanced that "Rather than viewing privacy concerns as impediment, policy makers, scientists and HIT specialists should embrace privacy as an opportunity that, if addressed, can enhance the flow of information" [19]. If patients' privacy is protected, this will ease and mitigate skepticism within those who are refusing to share their personal information for fear

that their privacy will be violated. These steps in the research process can facilitate the progression of scientific research through the increase of public participation and collaboration.

Besides addressing privacy concerns, researchers can focus on understanding what aspects of data need to be preserved and dispensed for the public good. As another potential barrier, data preservation and the awareness of what data needs to be preserved raises concerns about data quality, the absence of scientists to analyze data, and data storage options. Funding for research needs to be contingent upon the determination of the importance of digital data. Policies ought to be developed that relate to the use of data, such as what data to be preserved as well as what exceptions need to be made to data preservation. In addition, regulations about data hosts (warehouses for storing data) should be determined. For example, "Agencies and the research community together need to create the digital equivalent of libraries: institutions that can take responsibility for preserving digital data and making them accessible over the long term" [16]. Moreover, an effort to teach information management should be prioritized to facilitate data acquisition, data cleansing, data storage, and effective uses of data. While most scientific disciplines found that a data deluge is extremely challenging, great opportunities can be realized with better organization and open access to data [8]. It is important to train scientists, establish better policies to regulate data sharing, and increase the incentives for researchers from every fields to collaborate as they tackle the many issues that are faced by the modern sciences.

For data sharing and replication to be effective, scientists from diverse fields ought to come together because very rarely can progress happen in isolation. Currently, very few fields like astronomy, genomics, social sciences, and archaeology practice data sharing. The lack of success in implementing data sharing policies conveys the need for greater understanding of the roles of data in various sciences; highlighting the need to also seek the development of new models of scientific practice [6]. A new model can be in the arena of archiving; archiving data can be very expensive and difficult to manage, thus while it is encouraging that scientists share their work with each other, it is also crucial to have serious conversation about data housing, and the financial responsibility that involves in this practice. Until researchers come together to satisfy the response to those barriers, data sharing will remain a challenge among scientists. And if today's scientists and researchers do not make the effort to work together to facilitate effective and essential data sharing, future generations will experience the problem of lost data due to a lack of effective stewardship.

### 3.3 Ways to Overcome Privacy Concern

Three types of data are being identified: 1. personal and proprietary data, which are controlled by individuals and non-government organizations; 2. government controlled data, which includes, for instance, personal tax, census data, and personal health records; and finally, open data commons, which are available to everyone to access and use. The author advocated for policy makers develop strategy to link personal, proprietary, and government data to pursue health care care objectives [24]. When we think about privacy concerns it is crucial to see collaboration between scientists

from different sectors. By working together they will be more equipped to develop policies that can help to mitigate the risk of data leaking.

One way policymakers can protect individual privacy is by making the data anonymous. Researchers have identified three types of data: personal and proprietary data that is controlled by individuals; government-controlled data, which government agencies can restrict access to; and, open data commons, which means that the data is centrally located and available to all. Big data analysts and researchers have advocated for linking data together that can help to improve health care planning at both the patient and population levels. They also argued for an increase in the amount of information that is available in open data commons [17]. Although the anonymization of data appears to be a great technique that policymakers could espouse to address privacy concerns, other studies have indicated that some data can be traced back to their respective individual; thus, destroying the argument for anonymity [24]. "Every copy of data increases the risk of unintended disclosure. To reduce this risk, data should be anonymized before transfer; upon receipt, the recipient will have no choice but anonymize it at rest...And re-identification is by design, in order to ensure accountability, reconciliation and audit" [7] If proper norms are established for data analysis, this can potentially contribute to improvements in the health care industry, and businesses can maximize profit from it.

*3.3.1 Privacy Principles and Data Architecture.* Privacy principles should be introduced during the process of data architecture; privacy should be incorporated into the design and operational procedures [7]. In so doing, personal health care data, for example, will be protected against malicious hackers who try to access individuals' personal health information for the purposes of stealing individuals' identity. Another type of data that has been introduced to the healthcare industry is concept quantified self data. It can be understood as the data produced by individuals that engage in self-tracking of personal health information, such as heart rate, weight, energy levels, sleep quality, cognitive performance, etc. These individuals use devices like smart-phones, watches, and wearable technology sensors in the collection of their personal data and biometrics. It has been shown that 60 percent of U.S. adults are tracking their weight, diet or exercise routines, while 33 percent are monitoring their blood sugar, blood pressure, sleep patterns, etc. This indicates that there is a vast amount of health information that has been produced by individuals. What is done with all of this data? This massive supply demonstrates the need to develop policies and protocols that involve individual patient consent to share their collected data; this data can be critical to the advancement of health-care with the support of data analysis. Before that can be done; however, we must first establish the proper norm to use this type of data so that the privacy of individuals can be protected; this ought to be the primary action to take [22]. Because many individuals are willing to collect information about themselves without being prompted to do so; this is a good sign that is proper norm is established around data management and incentive is given to encourage data sharing, individuals will be willing to engage in the process of sharing their personal data. Although it is often

complicated to share qualitative data, the challenge to share seems to increase when dealing with healthcare data.

In the healthcare industry, Patients often do not want their health information to fall in the hand of other entities without their consent; however, with proper informed consent, patients seemed to become willing to share their personal health information. As agencies work with patients to disclose the purposes of collecting certain, sometimes sensitive, health information, they can empower patients to make informed decisions about their personal health information, thus engaging patients in the process. This can then serve to increase and improve the set of personal health information utilized for clinical research purposes, and subsequently improve people's lives [19]. "Privacy concerns exist wherever personally identifiable information or other sensitive information is collected and stored in any form" [12]. Thus, to protect privacy, other techniques, like encryption, authentication, and data masking may be utilized to ensure that the information is available only to authorized users.

## 4 SAVING SCIENTIFIC DATA FOR FUTURE GENERATIONS

### 4.1 Data Archiving

Along with the conversation surrounding data sharing and replication, another important conversation that needs to take place is around the infrastructure needed to archive data. While many people engage in a debate to encourage data sharing among scientists, the infrastructure to preserve the data does not yet fully exist. In reference to data, Nelson (2009) drew on a proverbial question, is it the chicken or the egg; what comes first, data sharing or the space to store data? He contends that while data sharing is encouraged among scientists, the infrastructure to store data is nonexistent and it is arduous task to pursue the development of it [15]. Thus, it is tantamount to talk about data saving as we encourage data sharing because if scientists resort to sharing their findings then there also need to be a safe place to house the data. This preservation is critical for future generations to build upon the work that has already been done to advance scientific research. As the advocacy for data archiving increases, a new challenge arises: who will pay for all of this?

**4.1.1 Who Should Pay to Store the Data?** Serious conversation needs to continue to happen around data management. "Access to data requires that the data be hosted somewhere and managed by someone" [4]. Although they acknowledged the effort of public and private sectors to archive data in certain fields like the life sciences, they also stipulated that many federally funded research data are at risk due to the lack of long-term structure that can ensure continual access and preservation of data. If data are not housed well, it could be said that a lot of efforts, money, and energies, are being wasted away due to lack of a secure and sustaining system of storage. As found in [14], the author posited that scientists are not the best stewards of data and suggested the job of data archiving be entrusted to the institutions that employ the researchers. Ensuring that data is well preserved will lay the important groundwork for allowing data to be accessible in the future. Lynch also emphasized the idea that for data saving to be effective, collaboration between funders, institutions and scientists are crucial. He gave the example,

such as the GenBank and the U.S.'s National Institutes of Health (NIH) genetic sequence database, as well as the U.S. National Virtual Observatory, to show the possibilities of what can be done. It appears that collaboration between sectors is key when it comes data management, including data archiving. It is not the job of one sector, but it is the job of all of us. Only when all of the sectors converge their effort together, we will be able to respond the quandary of Big data management.

Some researchers proposed four approaches that can help improve the partnership among sectors: 1) incentivize the private sector to be stewards of public research, 2) utilize the power of partnership between the public and private sectors to fund viable solutions, 3) create clear policies for the management of public data, and finally 4) encourage openness to new and diverse methods of research to advance public research abilities. Furthermore, they theorized that there should be adequate safeguards to prevent private sector's control, access and use of public data [4]. While these measures are applicable, there not be all great because by relinquishing the work of data saving to the private sector solely, it can create a very expensive problem to accessing data via private organizations that are highly incentivized by financial gains. It would probably be more effective if the federal government would pay for their own data scientists to be trained on how to effectively manage the data, and establish the requirements and expectations that all federally funded research remains in the public domain.

Other research corroborates the idea of licensing all research data to the public domain. This is the case, in the Netherlands, where for example, all the data retrievals are kept by the National Library. The U.S. can espouse this model by creating a center for data to be stored, and develop policies on how to access the data. This would prevent the private sector from having a monopoly on accessing, interpreting, sharing and store data that belongs in the public domain [18]. Another researcher advanced and noted that, in its effort to encourage peer review, the National Science Foundation (NSF), makes data sharing a requirement in the grant contract, where researchers are required to submit a 2-page report of their research that can be used to facilitate peer review. This technique used by NSF was revealed to be a great example of how the federal government can overcome the conundrum of data sharing and archiving among scientists [5]. This technique is applicable, however, proper norm stills to be established and proper security measure needs to be created to preclude the data leaking, which can compromise the privacy of research participants.

### 4.2 Electronic Versus Physical Archives

Another conversation that needs to happen when it comes to Big data management is how to archive the data. Scientists need to decide between electronic archiving versus physical archiving. "Digital archives face specific challenges linked to physical storage media as well as hardware and software longevity. In reality, every method of recording information, whether on paper, stone, or photographic film, has a limited resistance to time. The value of any information is dependent on the ability to decode it after a long storage period. The difference with digital archives is that these limits are ignored because of the addition of complex and versatile technology to the overall equation" [20]. Thus, it is important that researchers figure

out which form they want to espouse to archive that will allow long term access to the data. It can be argued that both forms have its advantages and disadvantages. Physical archiving allows access to the data locally and no sophisticated knowledge or programming skills are needed to access the data; however, the data is not available everywhere; this form of archiving data is quite limited. On the other hand, electronic archiving allows multiple users to access the data at the same time. It can be said that the digital information is not limited to physical location and space; the data is located in the cloud; it is unbound and can be accessed from everywhere. "The advantages of a digital archive in the pharmaceutical sector cover four areas: accessibility, selectivity, fidelity, and compliance" [20].

The only disadvantage to digital information is probably the lack of skills to access and manipulate the data. "As we move into the electronic era of digital objects, it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are, to my mind, living in the midst of digital Dark Ages; consequently, much as monks of times past, it falls to librarians and archivists to hold to the tradition which reveres history and the published heritage of our times" [20]. Thus, the necessity to continue to train more data scientists and analysts who can extract, manipulate, and analyze the information from the cloud system.

## 5 TRAIN NEW SCIENTIST TO HELP TO MANAGE DATA

If we are talking about data sharing, data replication and data archiving, it is critical to address the how to do so? The skills to extract, collect, and store data have not been taught yet in regular school. If we want to develop a generation of data driven society, we need to start teaching computer skills in elementary school all the way to college and university. Just like we teach natural language, like English, French, Spanish to children at a very young age, it is important to teach programming language skills to these kids, so that we can increase their awareness and develop incentive to become data analysts, scientists who will be able to manipulate the data sets. There is an old proverbial that states that " You can teach an old dog new tricks." Thus, the earlier we start teaching those skills, the better. While there are many institutions that have shown interest in developing Data Scientists by teaching programming language skills, like python, java, C++, machine learning, data analysis, etc. so that data can be extracted and analyzed, it would be great to start teaching programming skills to students at a lower level, like elementary school. The same way certain math skills like probability has been taught in high school, it is relevant to start incorporating programming language skills in high school to develop interest and incentive to engage in the process of collecting, manipulating, and housing digital information.

### 5.1 Why Data Repository?

Data repository facilitates access to existing documents. "Besides being a good thing for the sharing and verification of data-driven research results, data research repositories are now necessary for

university campuses. Placing one's research data online has become mandatory for any researcher wishing to receive grants from any public U.S. agency. This includes the National Institutes of Health (NIH), National Science Foundation (NSF), U.S. Department of Agriculture (USDA), and National Endowment for the Humanities (NEH). The rationale is that if a researcher is drawing from the public taxpayers' trough, the research must be publicly accessible through both the article and original data. Sharing this data helps keep the wider economy vital, facilitating healthy competition toward commercialization and dissemination of discovery. If researchers do not have data management plans in place, their chance of obtaining a grant decreases. Currently, a majority of grant-funded researchers do not share data. With recent mandated changes, this situation is rapidly changing. Ivy League institutions have already capitalized on it by sharing data leverages and enhances faculty, departmental, and a university's global research standing" [3]. If research data is made available, this can contribute to lessen and even mitigate the stress level of graduate students when they work on dissertation and final project for their respective institutions. Not only it would save them the time to go to different libraries to hunt for documents to start working on their project, but also it gives them access to all the work that has been done on the topic chosen, and what else is left to be done.

Data repositories can play a vital role in making research findings available to everyone and making access to data an easy process. "Online research data repositories are large database infrastructures set up to manage, share, access, and archive researchers' datasets. Repositories may be specialized and relegated to aggregating disciplinary data or more general, collecting over larger knowledge areas, such as the sciences or social sciences. Online repositories may also aggregate experts' data globally or locally, collecting a university or consortium of universities researcher's data for mutual benefit. The simple idea is that sharing data improves results and drives research and discovery forward. A repository allows examination, proof, review, transparency, and validation of a researcher's results by other experts beyond the published refereed academic article. Placing research data online allows instantaneous access by a globally dispersed group of researchers to share, understand, and synthesize results. This aggregation and synthesis provide an opportunity for insight, progress, and that uniquely human quest for larger understanding. Data repositories also allow for the publication of previously hidden negative data, essentially experiments that didn't work. This enables other researchers to avoid previous dead ends of those who have tried a path before them to better find their way toward more fertile territory" [3].

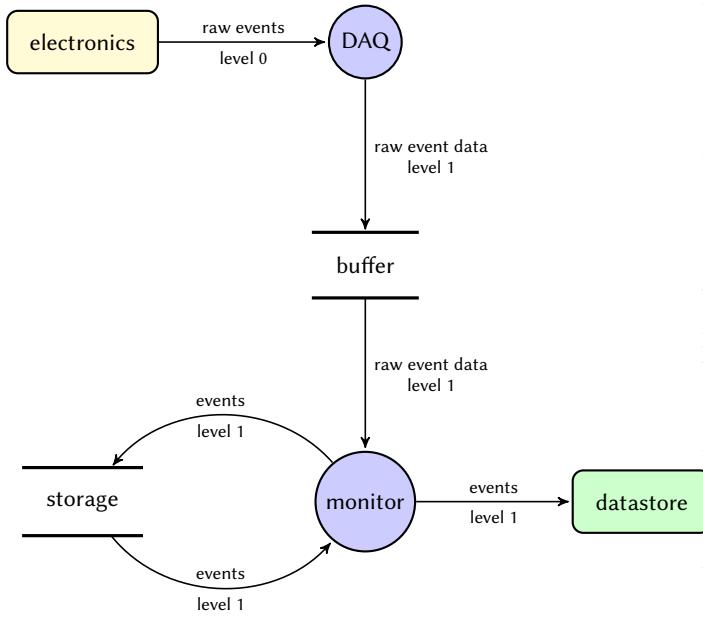
## 6 FIGURE

The following data flow diagrams convey the flow of information in a system. This figure shows experimental data being collected, reported, processed and stored [9]. This figure is depiction, an example of how data management works. The data go through a series of process before it can be archived and reused.

As we can see, managing big data requires that highly individuals to collect, clean, analyze, store, and share the data. It is not the work of one individual; it is the work of all of us. We need more data analysts who can engage in analyzing the data; we need more

data collectors to collect and create raw data and spreadsheets to be analyzed; we need more individuals who can create structures and security around data storing, etc. It was never the work of one individual, all of us need to be involved.

Thus, it is critical to converge our effort in passing the management tools and skills to current students and developing scientists who will be highly skilled in managing Big data analysis. Considering all of the different breakthroughs that are happening in the digital realm, it is safe to say that Big data is bigger and bigger; thus the necessity to train scientists to face the challenges that await us ahead.



## 6.1 figure2

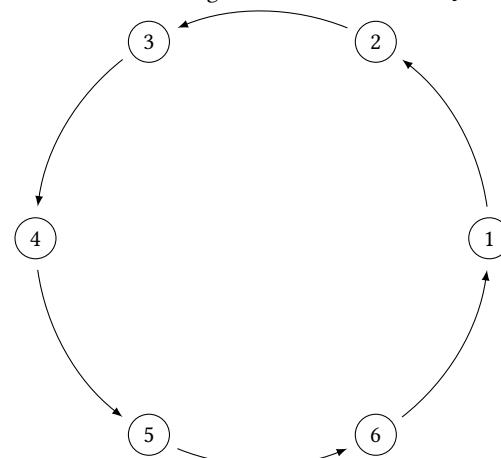
**6.1.1 A simple Example of Data lifecycle.** This data lifecycle was created by Jerome Tremblay and is adapted to explain the different phases that the data take before it can be destroyed [9].

The following figure is a depiction of the data lifecycle. The data often follows a number of steps, which are: in 1 we have (data creation), followed by 2 (data maintenance), 3 (data utilization), 4 (data publication), 5 (data archiving) and 6 (data destruction). In phase 1, the data gets created. Spreadsheets are often used by companies to keep the data. In this active process, the data is often stored locally on a server or multiple servers, or in the cloud, or a host data center. In phase 2, the data is data gets processed and synthesized in a variety of tasks. This is a fairly broad range of management actions, such as how data is supplied to the end users and how analytics such as modeling are performed on the data. In phase 3, the data is ready to be used by end users. In this phase the challenge of data governance and data compliance arise. In phase 4, the data can be published. In phase 5, the data is archive: At some point in time, the data in your system will have no immediate use, and it's time to archive it in case it might be needed in the future. This removes the data from your active environment and moves it off to storage. In phase 6, When the data is no longer

useful and needed, it must be destroyed. In this phase of the data lifecycle governance and compliance challenge might be surfaced. It's important to ensure that the data has actually been destroyed properly for several reasons, among those reason to make sure that privacy of individuals are protected. Briefly, these are the difference phases that data take before it falls into desuetude.

It is important to note that there are other types of data lifecycle in the realm of big data management that follow different stages, such as: data collection, in the this stage large amounts of raw data is being created; this stage is a significant aspect in the management of big data because it helps to capture the data that will later transition from raw data to published data. The second stage is data filtering and classification, in the stage the data is being filtered, cleaned and structured to eventually be ready to be analyzed. In the third stage, that data is ready to be analyzed. certain techniques and technologies are being used in the process of analyzing the data, such as data mining algorithm, cluster, correlation, statistical regression, indexing, graphics. Visualization and interpretation of the information happen in this very stage. The next stages that the data are storing, sharing, and publishing. After the data is being analyzed, the data is store for future use. "Data and its resources are collected and analyzed for storing, sharing, and publishing to benefit audiences, the public, tribal governments, academicians, researchers, scientific partners, federal agencies, and other stakeholders (e.g. industries, communities, and the media). Large and extensive Big data datasets must be stored and managed with reliability, availability, and easy accessibility, storage infrastructures must provide reliable space and a strong access interface that can not only analyze large amounts of data, but also store, manage, and determine data with relational DBMS structures. Storage capacity must be competitive given the sharp increase in data volume; hence, research on data storage is necessary" [11]. Thus, the importance to develop good policies that will address privacy and different challenges that involves storing and sharing Big data for the benefits of every sectors.

The codes for this figure was borrowed from Jerome Tremblay.



## 7 CONCLUSION

This document put forth the dialogue needed to assist researchers to address the different challenges they have experienced when

it comes to data sharing and replication as well as data saving. The advantages and disadvantages of Big data analysis have been discussed. We have seen that though Big data applications has its advantages, it has its poses many challenges as well. The importance of data sharing has been explored, and examples of how sharing information can help scientists to respond to global crises in a timely manner, like in the Ebola outbreak, have been provided. It has also been shown how data sharing and replication have helped to advance scientific research. It was postulated that in order for data to continue to exist, scientists need to embrace the idea of replicating their information.

As research continues to occur and scientists increase their agreement to collaborating with one another, they will be better equipped to discover potential errors from previous retrievals, fix those errors and clean the data, and make other discoveries based on existing data. Scientists cannot operate as an Island. Policies need to be put in place, to understand what data to share, when to share, where to store the data, and what data to store etc. For the continued advancement of the sciences, data sharing and archiving will require resources that facilitate the access, interpretation and maintenance of data. The importance of data sharing, data replication, and data archival cannot be overlooked. This work is far from exhaustive, More discussion around data management need to happen; new policies and regulations regarding how to share, store and replicate data are needed as well as effective parameters for how these processes will be funded and used in the future.

## ACKNOWLEDGMENTS

Thank you to Dr. Gregor von Laszewski for his support and suggestions to write this paper, and most importantly for teaching us how to use latex to write documents. This is a very important tool that everyone need to procure. It is very handy. From now on, latex is the new tool to write paper and article. I am so grateful for this class. Although I do not have any programming language background, being able to use latex to write documents is a big deal. I do not regret at all that I chose to take this course. It was a pleasure to be in this class. I have learned a lot in the course. I will definitely recommend this course to other students. Despite my meager python and programming skill, with the assistance of the TAs and the professor, I was able to respond to the challenge of this class. Thus, thank you so much everyone for your help and assistance .

## REFERENCES

- [1] [n. d.]. ([n. d.]). <https://www.encyclopedia2.thefreedictionary.com/data+sharing>
- [2] [n. d.]. ([n. d.]). <http://www.searchdisasterrecovery.techtarget.com/definition/data-replication>
- [3] [n. d.]. ([n. d.]). <http://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml>
- [4] Francine Berman and Vint Cerf. 2013. Who will pay for public access to research data? *Science* 341, 6146 (2013), 616–617.
- [5] Christine L Borgman. 2012. The conundrum of sharing research data. *Journal of the Association for Information Science and Technology* 63, 6 (2012), 1059–1078.
- [6] Christine L Borgman. 2015. If data sharing is the answer, what is the question? *ERCIM NEWS* (2015), 15.
- [7] Ann Cavoukian and Jeff Jonas. 2012. *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada.
- [8] TO SPUR ECONOMIC. 2011. Challenges and Opportunities. *databases* 22 (2011), 21–4.
- [9] D. Fokkema and D. Fokkema. 2012. The Hisparc cosmic ray experiment : data acquisition and reconstruction of shower direction. (10 2012). <https://doi.org/10.3990/1.9789036534383>
- [10] Richa Gupta, Sunny Gupta, and Anuradha Singhal. 2014. Big data: overview. *arXiv preprint arXiv:1404.4136* (2014).
- [11] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal* 2014 (2014), 18.
- [12] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. 2016. Digital Health Data: A Comprehensive Review of Privacy and Security Risks and Some Recommendations. *Computer Science Journal of Moldova* 24, 2 (2016), 71.
- [13] Kalev Leetaru. 2016. Data Sharing and Replication in the Sciences. *Science* (2016).
- [14] Clifford Lynch. 2008. Big data: How do your data grow? *Nature* 455, 7209 (2008), 28–29.
- [15] Bryn Nelson. 2009. Empty archives: most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? Bryn Nelson investigates why many researchers choose not to share. *Nature* 461, 7261 (2009), 160–164.
- [16] Graham Pryor and Martin Donnelly. 2009. Skilling up to do data: whose role, whose responsibility, whose career? *International Journal of Digital Curation* 4, 2 (2009), 158–170.
- [17] Joachim Roski, George W Bo-Linn, and Timothy A Andrews. 2014. Creating value in health care through big data: opportunities and policy implications. *Health affairs* 33, 7 (2014), 1115–1122.
- [18] Vera Sarkol. 2016. Scientific data and preservation-policy issues for the long-term record. *ERCIM News* 107 (2016), 13–14.
- [19] Robert H Shelton. 2011. Electronic consent channels: preserving patient privacy without handcuffing researchers. *Science translational medicine* 3, 69 (2011), 69cm4–69cm4.
- [20] Dimitri Stamatiadis. 2005. Digital archiving in the pharmaceutical industry. *Information Management* 39, 4 (2005), 54.
- [21] Mark Stoakes and Katherine Irwin. 2005. Data Replication and Data Sharing—Integrating Heterogeneous Spatial Databases. (2005), 12 pages.
- [22] Melanie Swan. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data* 1, 2 (2013), 85–99.
- [23] Omer Tene and Jules Polonetsky. 2012. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11 (2012), xxvii.
- [24] J Van Den Bos, K Rustagi, T Gray, M Halford, E Zeimkiewicz, and J Shreve. 2011. Health affairs: At the intersection of health, health care and policy. *Health Affairs* 30 (2011), 596–603.
- [25] Gretchen Vogel. 2014. Delays hinder Ebola genomics. *Science* 346, 6210 (2014), 684–685.
- [26] Nathan L Yozwiak, Stephen F Schaffner, and Pardis C Sabeti. 2015. Data sharing: Make outbreak research open access. *Nature News* 518, 7540 (2015), 477.

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--no key, author, or editor in Datasharing  
Warning--no author, editor, organization, or key in Datasharing  
Warning--to sort, need author, editor, or key in Datasharing  
Warning--no key, author, or editor in Datarep2017  
Warning--no author, editor, organization, or key in Datarep2017  
Warning--to sort, need author, editor, or key in Datarep2017  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no author, editor, organization, or key in Uswyshyn2016  
Warning--to sort, need author, editor, or key in Uswyshyn2016  
Warning--no key, author, or editor in Datasharing  
Warning--no key, author, or editor in Datasharing  
Warning--no key, author, or editor in Datarep2017  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no key, author, or editor in Datasharing  
Warning--neither author and editor supplied for Datasharing  
Warning--empty year in Datasharing  
Warning--empty title in Datasharing  
Warning--no journal in Datasharing  
Warning--no number and no volume in Datasharing  
Warning--page numbers missing in both pages and numpages fields in Datasharing  
Warning--no key, author, or editor in Datarep2017  
Warning--no author, editor, organization, or key in Datarep2017  
Warning--neither author and editor supplied for Datarep2017  
Warning--empty year in Datarep2017  
Warning--empty title in Datarep2017  
Warning--no journal in Datarep2017  
Warning--no number and no volume in Datarep2017  
Warning--page numbers missing in both pages and numpages fields in Datarep2017  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no author, editor, organization, or key in Uswyshyn2016  
Warning--neither author and editor supplied for Uswyshyn2016  
Warning--empty year in Uswyshyn2016  
Warning--empty title in Uswyshyn2016  
Warning--no journal in Uswyshyn2016  
Warning--no number and no volume in Uswyshyn2016  
Warning--page numbers missing in both pages and numpages fields in Uswyshyn2016

```
Warning--no number and no volume in borgman2015if
Warning--no number and no volume in gupta2014big
Warning--page numbers missing in both pages and numpages fields in gupta2014big
Warning--no number and no volume in leetaru2016
(There were 42 warnings)
```

bibtext \_ label error

---

bibtext space label error  
=====

report.bib:209:@ Article{Datasharing,

bibtext comma label error

latex report

[2017-12-16 09.37.39] pdflatex report.tex

```
=====
Compliance Report
=====
```

```
name: J. Robert Langlois
hid: 325
paper1: Oct 23 17 100%
paper2: 100%
project: 100%
```

```
yamlcheck
-----
```

```
wordcount
-----
```

```
8
wc 325 project 8 7721 report.tex
wc 325 project 8 7758 report.pdf
wc 325 project 8 889 report.bib
```

```
find "
-----
```

78: Nowadays, many fields are witnessing a large influx of data due to the increase usage of technology. The digital data generated from scientific research is integral to the advancement of different scientific fields. While these data sets exist in abundant quantities, one challenge that many fields face is the lack of and/or prohibition of data being shared among researchers. Data sharing and its subsequent replication is a subject matter that is in dispute within in the sciences \cite{leetaru2016}. This is a significant area of contention in the United States; however, in other countries like the United Kingdom (U.K.), this issue has been addressed by making data sharing a matter of great importance; so much so, that the Joint Information Systems Committee of the U.K. made "data-sharing a priority, and has helped to establish a Digital Curation Centre to be national focus for research and development into data issues" \cite{pryor2009skilling}. On one hand, opponents of data sharing are skeptical about this practice due to privacy concerns, fears of stigmatization, funding problems, repositories for data,

transparency, etc. On the other hand, some researchers are open to data sharing because it allows their work to be reviewed and creates opportunities to further their findings; however, the actual practice is stunted by researchers concerns

\cite{nelson2009empty}. Proponents of data sharing continue to advocate for an open access policy that would allow data to quickly respond to societal problems and crises, as well as advance the sciences. While it is as important to keep in mind the different challenges to data sharing, like data archiving, sharing data among fellow scientists can be very beneficial, not only can this practice help to maximize profits, make new discoveries, and respond to crises more quickly, but also it can play a vital role in advancing science and research.

- 88: Big data "is data that exceeds the processing capacity of traditional databases. The data is too big to be processed by a single machine. New and innovative methods are required to process and store such large volumes of data" \cite{gupta2014big}. The data sets are so voluminous and complex that traditional way of data processing application software are insufficient to deal with them.
- 92: Data replication is the process of copying data from one location to another. The technology helps an organization to possess up-to-date copies of its data in the event of a disaster."Replication can take place over a storage area network, local area network or local wide area network, as well as to the cloud. For disaster recovery (DR) purposes, replication typically occurs between a primary storage location and a secondary offsite location" \cite{Datarep2017}.
- 111: Certain research studies have supported the idea that big data allows for real time tracking of diseases and the development, prediction of outbreaks, and facilitates the development of personalized healthcare. Big data can also be used to maximize profits in many disciplines, including healthcare if harnessed properly \cite{van2011health}. "By harnessing big data, businesses gain many advantages, including increased operational efficiency, informed strategic direction, improved customer service, new products, and new customers and markets" \cite{khan2014big}. While data exists in huge quantities in many fields, including the health care field, individual privacy concerns remain a big problem that policymakers have to tackle to meet current trends in data collection. Improved methods of protecting very personal, private and sensitive health information is needed in order to allow for safe, necessary and

adequate access to protected health information within the health care industry. Without proper policies related to data use, access, and protection, this big data potential can not be realized \cite{roski2014creating}. Data sharing and replication contribute to increase the availability of the amounts of digital information and make the access to information easier.

- 116: Furthermore, data sharing and replication (the availability of multiple copies of a data set to different users) is an important practice that plays a crucial role in the advancing of the sciences. The way sciences grow is through scaffolding, which means that one must rely on the work of previously published research to come up with new findings; thus, further supporting the necessity for a collaborative effort among researchers and scientists. Data sharing has been shown to help spot errors in research. In 2013, for example, a graduate student pointed out a calculation error made by two Harvard professors. This discovery was only possible because the professors shared a spreadsheet of their research findings with the particular student \cite{leetaru2016}. In this case, and possibly in many other cases, data sharing has helped to build community among fellow researchers, uncover honest mistakes and, in worst-case scenarios, expose possible fraud. Another researcher made a series of observations about the relevance of sharing data and asked many poignant questions regarding content, parameters and the necessity for sharing information. One observation she made was that "science progressed for centuries without data sharing policies and then questioned, why is data sharing deemed so important to scientific progress now?" \cite{borgman2015if}. She challenged the notion of free and unhindered distribution of data and cautioned that preliminary questions must first be answered to determine what data to share, how much and in what context data should be released to advance the changes in sciences. Her point was that the data should not stand alone, but rather it ought to be accurately defined and contextualized within the means that identified, developed and synthesized the data into usable information. While Borgman's scrutiny has its place in the argument regarding the absence of policies which she argues ought to be based on accurately defining the data parameters and their ability to facilitate data sharing, it is also important to acknowledge that the various scientific fields have been faced with different types of data and challenges.
- 124: Another reason to encourage data sharing is that data sharing help to reduce unnecessary cost. As digital information is made available in the cloud system, researchers will be able to access

the same type of information at different locations. "Data sharing is driven by the need to maintain more accurate and up-to-date spatial databases, but at the same time reduce data acquisition and maintenance costs" \cite{stoakes2005data}. If data becomes more accessible, not only this will contribute to lowering the cost to access data, but also will it encourage researchers in their endeavors to respond to social outbreak quicker. Data sharing can also play a crucial role in advancing science, which is our next point.

- 128: Contrarily to what many might think, the advancement science occurs through replication of existing findings; scientists must rely on the work of previous researchers to make new discoveries. Just like human being cannot live in vacuum; the way science develop is through collaborative effort among fellow scientists."Placing research data online allows instantaneous access by a globally dispersed group of researchers to share, understand, and synthesize results. This aggregation and synthesis provide an opportunity for insight, progress, and that uniquely human quest for larger understanding. Data repositories also allow for the publication of previously hidden negative data, essentially experiments that didn't work" \cite{Uswyshyn2016}. One advantage of this practice is that by sharing their work, scientists will be able to spout errors that previous researchers have made, reveal fraud, build community, etc \cite{leetaru2016}. As found in \cite{borgman2012conundrum} policy makers must develop policies that explain how to embark in this process. If data are not being replicated, the world of science will face far more challenges in the future. Data replication involves open access to data so that researchers can continue to study, analyze, and make new conclusion about existing data. Now, if data sharing allows the replication of the sciences, why are many people opposed such practice?
- 138: Some researchers fear that their work is going to be poached and that they will not get credit for their findings, so they hold on to it and do not disclose it. The concerns of obscurity and/or credit being assigned to other researchers who might advance the original researcher's findings will cause a serious reluctance to sharing data. As found in \cite{leetaru2016}, the term data parasites is used to describe the practice of utilizing data without giving proper credit to prior publishers. Thus, to overcome this challenge, there should be honesty and allowance for intellectual probity; a sort of fair play between researchers where appropriate credit would be given. In addition to giving appropriate credit for findings, another important aspect of

disclosure is the appropriate compensation for the release of intellectual property. Oftentimes, researcher have sacrificed their time, income potential, energies, and relationships to do the work they do; incentive needs to be provided to encourage the sharing of their findings that they have worked hard to develop. "The call for transparency is not new, of course. Rather the emphasis is on access to data in a usable format, which can work to create value to individuals" \cite{tene2012big}. Access to digital information can engage individuals, invite scrutiny, and expose misuses of data.

- 145: Another barrier to data sharing, specifically in the healthcare field, involves the protection of patient privacy; a lack thereof can lead to stigmatization and potentially hamper patients participation in healthcare research and treatment. "Privacy is a major concern in outsourced data. recently some controversies have revealed how some security agencies are using data generated by individuals for their own benefits without permission. Therefore, policies that cover all user privacy concerns should be developed. Furthermore, rule violators should be identified and users data should not be misused or leaked" \cite{khan2014big}. In so doing, individuals will feel more at ease to engage in the process of sharing information for the benefits of everyone.
- 148: As \cite{yozwiak2015data} highlighted, some uncertainties that are involved data sharing, like whether data belongs to public or private domains. Still, another barrier is patient consent and their ability to fully understand how their participation can make them vulnerable to being potentially shunned and ostracized in their community based on their diagnosis and/or treatment. The researchers advocated for the responsible sharing of pertinent information among researchers to avoid this problem. It should also be mentioned that preclusion to the sharing of unnecessary information would also weaken the barriers to data sharing. Although data sharing is important, particularly during a medical outbreak, researchers ought to do their best to protect patient privacy to avoid any threat of stigmatization or isolation of patients. Rigorous ethical standards should be applied to safeguard patients' privacy and dignity to allow for easier sharing of relevant data \cite{yozwiak2015data}. Shelton (2011) advanced that "Rather than viewing privacy concerns as impediment, policy makers, scientists and HIT specialists should embrace privacy as an opportunity that, if addressed, can enhance the flow of information" \cite{shelton2011electronic}. If patients' privacy is protected, this will ease and mitigate

skepticism within those who are refusing to share their personal information for fear that their privacy will be violated. These steps in the research process can facilitate the progression of scientific research through the increase of public participation and collaboration.

- 151: Besides addressing privacy concerns, researchers can focus on understanding what aspects of data need to be preserved and dispensed for the public good. As another potential barrier, data preservation and the awareness of what data needs to be preserved raises concerns about data quality, the absence of scientists to analyze data, and data storage options. Funding for research needs to be contingent upon the determination of the importance of digital data. Policies ought to be developed that relate to the use of data, such as what data to be preserved as well as what exceptions need to be made to data preservation. In addition, regulations about data hosts (warehouses for storing data) should be determined. For example, "Agencies and the research community together need to create the digital equivalent of libraries: institutions that can take responsibility for preserving digital data and making them accessible over the long term" \cite{pryor2009skilling}. Moreover, an effort to teach information management should be prioritized to facilitate data acquisition, data cleansing, data storage, and effective uses of data. While most scientific disciplines found that a data deluge is extremely challenging, great opportunities can be realized with better organization and open access to data \cite{economic22challenges}. It is important to train scientists, establish better policies to regulate data sharing, and increase the incentives for researchers from every fields to collaborate as they tackle the many issues that are faced by the modern sciences.
- 162: One way policymakers can protect individual privacy is by making the data anonymous. Researchers have identified three types of data: personal and proprietary data that is controlled by individuals; government-controlled data, which government agencies can restrict access to; and, open data commons, which means that the data is centrally located and available to all. Big data analysts and researchers have advocated for linking data together that can help to improve health care planning at both the patient and population levels. They also argued for an increase in the amount of information that is available in open data commons \cite{roski2014creating}. Although the anonymization of data appears to be a great technique that policymakers could espouse to address privacy concerns, other studies have indicated

that some data can be traced back to their respective individual; thus, destroying the argument for anonymity \cite{van2011health}. "Every copy of data increases the risk of unintended disclosure. To reduce this risk, data should be anonymized before transfer; upon receipt, the recipient will have no choice but anonymize it at rest...And re-identification is by design, in order to ensure accountability, reconciliation and audit"

\cite{cavoukian2012privacy} If proper norms are established for data analysis, this can potentially contribute to improvements in the health care industry, and businesses can maximize profit from it.

- 171: "Privacy concerns exist wherever personally identifiable information or other sensitive information is collected and stored in any form" \cite{khan2016digital}. Thus, to protect privacy, other techniques, like encryption, authentication, and data masking may be utilized to ensure that the information is available only to authorized users.
- 184: Serious conversation needs to continue to happen around data management. "Access to data requires that the data be hosted somewhere and managed by someone" \cite{berman2013will}. Although they acknowledged the effort of public and private sectors to archive data in certain fields like the life sciences, they also stipulated that many federally funded research data are at risk due to the lack of long-term structure that can ensure continual access and preservation of data. If data are not housed well, it could be said that a lot of efforts, money, and energies, are being wasted away due to lack of a secure and sustaining system of storage. As found in \cite{lynch2008big}, the author posited that scientists are not the best stewards of data and suggested the job of data archiving be entrusted to the institutions that employ the researchers. Ensuring that data is well preserved will lay the important groundwork for allowing data to be accessible in the future. Lynch also emphasized the idea that for data saving to be effective, collaboration between funders, institutions and scientists are crucial. He gave the example, such as the GenBank and the U.S.'s National Institutes of Health (NIH) genetic sequence database, as well as the U.S. National Virtual Observatory, to show the possibilities of what can be done. It appears that collaboration between sectors is key when it comes data management, including data archiving. It is not the job of one sector, but it is the job of all of us. Only when all of the sectors converge their effort together, we will be able to respond the quandary of Big data management.

195: Another conversation that needs to happen when it comes to Big data management is how to archive the data. Scientists need to decide between electronic archiving versus physical archiving. "Digital archives face specific challenges linked to physical storage media as well as hardware and software longevity. In reality, every method of recording information, whether on paper, stone, or photographic film, has a limited resistance to time. The value of any information is dependent on the ability to decode it after a long storage period. The difference with digital archives is that these limits are ignored because of the addition of complex and versatile technology to the overall equation" \cite{stamatiadis2005digital}. Thus, it is important that researchers figure out which form they want to espouse to archive that will allow long term access to the data. It can be argued that both forms have its advantages and disadvantages. Physical archiving allows access to the data locally and no sophisticated knowledge or programming skills are needed to access the data; however, the data is not available everywhere; this form of archiving data is quite limited. On the other hand, electronic archiving allows multiple users to access the data at the same time. It can be said that the digital information is not limited to physical location and space; the data is located in the cloud; it is unbound and can be accessed from everywhere. "The advantages of a digital archive in the pharmaceutical sector cover four areas: accessibility, selectivity, fidelity, and compliance" \cite{stamatiadis2005digital}.

197: The only disadvantage to digital information is probable the lack of skills to access and manipulate the data. "As we move into the electronic era of digital objects, it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are, to my mind, living in the midst of digital Dark Ages; consequently, much as monks of times past, it falls to librarians and archivists to hold to the tradition which reveres history and the published heritage of our times" \cite{stamatiadis2005digital}. Thus, the necessity to continue to train more data scientists and analysts who can extract, manipulate, and analyze the information from the cloud system.

201: If we are talking about data sharing, data replication and data archiving, it is critical to address the how to do so? The skills to extract, collect, and store data have not been taught yet in regular school. If we want to develop a generation of data driven society, we need to start teaching computer skills in elementary

school all the way to college and university. Just like we teach natural language, like English, French, Spanish to children at a very young age, it is important to teach programming language skills to these kids, so that we can increase their awareness and develop incentive to become data analysts, scientists who will be able to manipulate the data sets. There is an old proverbial that states that " You can teach an old dog new tricks." Thus, the earlier we start teaching those skills, the better. While there are many institutions that have shown interest in developing Data Scientists by teaching programming language skills, like python, java, C++, machine learning, data analysis, etc. so that data can be extracted and analyzed, it would be great to start teaching programming skills to students at a lower level, like elementary school. The same way certain math skills like probability has been taught in high school, it is relevant to start incorporating programming language skills in high school to develop interest and incentive to engage in the process of collecting, manipulating, and housing digital information.

207: Data repository facilitates access to existing documents.

"Besides being a good thing for the sharing and verification of data-driven research results, data research repositories are now necessary for university campuses. Placing ones research data online has become mandatory for any researcher wishing to receive grants from any public U.S. agency. This includes the National Institutes of Health (NIH), National Science Foundation (NSF), U.S. Department of Agriculture (USDA), and National Endowment for the Humanities (NEH). The rational is that if a researcher is drawing from the public taxpayers trough, the research must be publicly accessible through both the article and original data. Sharing this data helps keep the wider economy vital, facilitating healthy competition toward commercialization and dissemination of discovery. If researchers do not have data management plans in place, their chance of obtaining a grant decreases. Currently, a majority of grant-funded researchers do not share data. With recent mandated changes, this situation is rapidly changing. Ivy League institutions have already capitalized on its sharing data leverages and enhances faculty, departmental, and a university's global research standing" \cite{Uswyshyn2016}. If research data is made available, this can contribute to lessen and even mitigate the stress level of graduate students when they work on dissertation and final project for their respective institutions. Not only it would save them the time to go to different libraries to hunt for documents to start working on their project, but also it gives them access to all the work that has been done on the topic chosen, and what

else is left to be done.

- 210: Data repositories can play a vital role in making research findings available to everyone and making access to data an easy process. "Online research data repositories are large database infrastructures set up to manage, share, access, and archive researchers' datasets. Repositories may be specialized and relegated to aggregating disciplinary data or more general, collecting over larger knowledge areas, such as the sciences or social sciences. Online repositories may also aggregate experts' data globally or locally, collecting a university or consortium of universities researcher's data for mutual benefit. The simple idea is that sharing data improves results and drives research and discovery forward. A repository allows examination, proof, review, transparency, and validation of a researcher's results by other experts beyond the published refereed academic article. Placing research data online allows instantaneous access by a globally dispersed group of researchers to share, understand, and synthesize results. This aggregation and synthesis provide an opportunity for insight, progress, and that uniquely human quest for larger understanding. Data repositories also allow for the publication of previously hidden negative data, essentially experiments that didn't work. This enables other researchers to avoid previous dead ends of those who have tried a path before them to better find their way toward more fertile territory"  
\cite{Uswyshyn2016}.
- 273: It is important to note that there are other types of data lifecycle in the realm of big data management that follow different stages, such as: data collection, in the this stage large amounts of raw data is being created;this stage is a significant aspect in the management of big data because it helps to capture the data that will later transition from raw data to published data. The second stage is data filtering and classification, in the stage the data is being filtered, cleaned and structured to eventually be ready to be analyzed. In the third stage, that data is ready to be analyzed. certain techniques and technologies are being used in the process of analyzing the data, such as data mining algorithm, cluster, correlation, statistical regression, indexing, graphics. Visualization and interpretation of the information happen in this very stage. The next stages that the data are storing, sharing, and publishing. After the data is being analyzed, the data is store for future use. "Data and its resources are collected and analyzed for storing, sharing, and publishing to benefit audiences, the public, tribal governments, academicians,

researchers, scientific partners, federal agencies, and other stakeholders (e.g. industries, communities, and the media). Large and extensive Big data datasets must be stored and managed with reliability, availability, and easy accessibility, storage infrastructures must provide reliable space and a strong access interface that can not only analyze large amounts of data, but also store, manage, and determine data with relational DBMS structures. Storage capacity must be competitive given the sharp increase in data volume; hence, research on data storage is necessary" \cite{khan2014big}. Thus, the importance to develop good policies that will address privacy and different challenges that involves storing and sharing Big data for the benefits of every sectors.

passed: False

find footnote

---

234: to/.style={->, >=stealth', shorten  
>=1pt, semithick, font=\sffamily\footnotesize},

passed: False

find input{format/i523}

---

37: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

figures 0  
tables 0  
includegraphics 0  
labels 0  
refs 0  
floats 0

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
True : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

---

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no key, author, or editor in Datasharing
Warning--no author, editor, organization, or key in Datasharing
Warning--to sort, need author, editor, or key in Datasharing
Warning--no key, author, or editor in Datarep2017
Warning--no author, editor, organization, or key in Datarep2017
Warning--to sort, need author, editor, or key in Datarep2017
Warning--no key, author, or editor in Uswyshyn2016
Warning--no author, editor, organization, or key in Uswyshyn2016
Warning--to sort, need author, editor, or key in Uswyshyn2016
Warning--no key, author, or editor in Datasharing
```

Warning--no key, author, or editor in Datasharing  
Warning--no key, author, or editor in Datarep2017  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no key, author, or editor in Datasharing  
Warning--no author, editor, organization, or key in Datasharing  
Warning--neither author and editor supplied for Datasharing  
Warning--empty year in Datasharing  
Warning--empty title in Datasharing  
Warning--no journal in Datasharing  
Warning--no number and no volume in Datasharing  
Warning--page numbers missing in both pages and numpages fields in Datasharing  
Warning--no key, author, or editor in Datarep2017  
Warning--no author, editor, organization, or key in Datarep2017  
Warning--neither author and editor supplied for Datarep2017  
Warning--empty year in Datarep2017  
Warning--empty title in Datarep2017  
Warning--no journal in Datarep2017  
Warning--no number and no volume in Datarep2017  
Warning--page numbers missing in both pages and numpages fields in Datarep2017  
Warning--no key, author, or editor in Uswyshyn2016  
Warning--no author, editor, organization, or key in Uswyshyn2016  
Warning--neither author and editor supplied for Uswyshyn2016  
Warning--empty year in Uswyshyn2016  
Warning--empty title in Uswyshyn2016  
Warning--no journal in Uswyshyn2016  
Warning--no number and no volume in Uswyshyn2016  
Warning--page numbers missing in both pages and numpages fields in Uswyshyn2016  
Warning--no number and no volume in borgman2015if  
Warning--no number and no volume in gupta2014big  
Warning--page numbers missing in both pages and numpages fields in gupta2014big  
Warning--no number and no volume in leetaru2016  
(There were 42 warnings)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
non ascii found 8230
non ascii found 8212
non ascii found 8212
non ascii found 8217
non ascii found 8217
non ascii found 8217
non ascii found 8217
non ascii found 8212
non ascii found 8217
non ascii found 8217
non ascii found 8217
```

---

The following tests are optional

---

Tip: newlines can often be replaced just by an empty line

find newline

---

```
passed: True
cites should have a space before \cite{} but not before the {
```

find cite {

---

passed: True

# Continuous Motion Detection Using Convolutional Neural Network and Recurrent Neural Network

Ajinkya Khamkar

Indiana University

P.O. Box 1212

Bloomington, Indiana 47408

adkhamka@iu.edu

## ABSTRACT

Object detection is fundamental and important Computer Vision task. Continuous object detection is an extension of object detection for continuous motion scenes. Traditional methodologies include estimating object displacement in subsequent scenes using optical flow methodology. Optical flow can be computationally expensive to compute making it unattractive for online learning methodologies. Deep Neural Network learning techniques present an alternative approach for determining continuous motion without explicitly computing the optical flow between subsequent frames.

## KEYWORDS

I523,HID211, Continuous Motion Detection, Convolutional Neural Networks, Object Detection, Deep Neural Networks

## 1 INTRODUCTION

Object detection is fundamental and important Computer Vision task. Continuous object detection is an extension of object detection for continuous motion scenes. In section 2, we discuss the scope and applications driven by continuous motion detection. In section 3, we present the data that we use for our experiments.

In section 4, we discuss traditional techniques used for continuous motion detection. Traditionally, hand crafted features [14] and optical flow [13] between subsequent frames was used to detect continuous motion detection. We discuss the major drawbacks of using traditional optical flow and hand crafted feature based methods. These drawbacks can be overcome using newer deep learning techniques.

We begin section 5 by discussing about deep convolutional neural networks and their application for object detection. We introduce certain naive methods that can be used to achieve continuous object detection. In section 5.3, we introduce Recurrent Neural Networks and their application for continuous object detection. We discuss long-short-term-memory networks a form of recurrent neural network which is designed to retain scene memory over long periods of time. We discuss implementations which use long-short-term-memory networks along with deep convolutional neural networks to improve the performance of the naive methods.

In section 6, we present an end-to-end approach and algorithm which can be trained in a single shot fashion with the gradient generated by long-short-term-memory network to train the object detection network.

In section 6.1, we discuss the our training methodology and training resources used for our experiment. In section 7, we present

the results we achieved for training the model in an end-to-end fashion. In section 8, we conclude our discussion.

## 2 APPLICATIONS

The applications and scope of motion detection and object tracking has risen exponentially over the last decade. In recent years, we have seen several concepts of automated vehicular driving, physical robots aiding human-centric activities, aerial drone technologies replacing traditional delivery schemes, virtual reality based systems tracking human movement habits and learning from those and use of motion tracking in popular sports for crucial decision making[8]. Continuous motion detection and robotic vision remain at the crux of all the above applications. Traditionally, motion detection has also been used for security monitoring [1] and tracking suspicious activities. Researchers have traditionally used object tracking methodologies to study human movement patterns, track bird, mammal and aquatic migration patterns and draw conclusions from them. Thus it remains important to introduce computationally efficient and online learning methodologies which can be used for real time object motion detection.

## 3 DATA

One of the major difficulties in training architectures for continuous motion detection is lack of availability of labelled data. Videos are a sequence of image frames and speed of motion is determined using the rate at which these frames are presented to the naked human eye. Additionally motion changes in subsequent frames is minimal and can be approximated to zero or no-motion. Human experts are required to annotate images by drawing bounding boxes around objects of interests in image frames. Few second long videos can have thousands of frames depending upon the frame rate. Several online platforms including amazons mechanical turk are used by researchers to create labelled data. Participants are paid for labelling scenes and hand annotating object locations. This makes the annotation task expensive and tedious.

For this experiment we use Visual Object Tracking challenge dataset [5]. The dataset has 16 videos corresponding to different action sequences, each action has on an average 400 motion frames. Each frame is further labelled with the object of interest and bounding box annotations. Each frame presents a single object of interest. Each object is under varied illumination, colour and shape conditions. Additionally the images are noisy and blurry replicating low resolution tracking devices traditionally found in the wild.

## 4 TRADITIONAL APPROACHES

Traditional methods involve use of hand crafted features including hand annotating interest points on objects within images and then tracking the motion of interest points in subsequent frames. Another traditional approach uses the bag of words representation of images or uses the histogram of oriented gradients [14] approach to locate objects of interest in the images. These approaches have multiple severe drawbacks which we discuss further.

- These approaches fail to scale for larger dataset with different objects of varied size and shapes. They suffer further from scale and illumination variance in continuous frames.
- Researchers are tasked with detecting feature points within images, thus making this mechanically tasking and expensive.

Another important and popular approach for tackling continuous motion problems involves the computation of optical flow. Optical flow [13] computes the relative motion between pixels in subsequent frames. Optical flow uses the following assumption, given motion of the pixels in subsequent frames is small or negligible, the changes in intensity or brightness in subsequent frames will be constant or near zero.

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (1)$$

In equation 1  $I$  represents the pixel wise intensity across frames. If the movement is small the right hand side of the above equation can be approximated using the first order Taylor series

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\delta I}{\delta x} \Delta x + \frac{\delta I}{\delta y} \Delta y + \frac{\delta I}{\delta t} \Delta t \quad (2)$$

$$\frac{\delta I}{\delta x} \Delta x + \frac{\delta I}{\delta y} \Delta y + \frac{\delta I}{\delta t} \Delta t = 0 \quad (3)$$

$$\frac{\delta I}{\delta x} \frac{\Delta x}{\Delta t} + \frac{\delta I}{\delta y} \frac{\Delta y}{\Delta t} + \frac{\delta I}{\delta t} \frac{\Delta t}{\Delta t} = 0 \quad (4)$$

$$\frac{\delta I}{\delta x} \frac{\Delta x}{\Delta t} + \frac{\delta I}{\delta y} \frac{\Delta y}{\Delta t} + \frac{\delta I}{\delta t} = 0 \quad (5)$$

$$\frac{\delta I}{\delta x} V_x + \frac{\delta I}{\delta y} V_y = -\frac{\delta I}{\delta t} \quad (6)$$

Here  $V_x$  and  $V_y$  represent the optical flow in directions  $x, y$  and the partial derivative represent the derivative of the image at pixel  $x$  and  $y$ . The equation has 2 unknowns and requires additional constraints and equation to solve for the unknowns. Two popular approaches to estimate the optical flow include Lucas-Kanade method and Horn-Schunk method. Lucas-Kanade [13] method solves for the unknown using the assumption, the motion for all pixels in a window centered around  $p$  between frames will be constant. Thus they solve for the equation.

$$\begin{aligned} I_x(p_1)V_x + I_y(p_1)V_y &= -I_t(p_1) \\ I_x(p_2)V_x + I_y(p_2)V_y &= -I_t(p_2) \\ &\vdots \end{aligned} \quad (7)$$

Here  $I_{x,y}(p_n)$  represents the partial derivative of the intensity at location  $x, y$  and  $V_{x,y}$  is the optical flow at pixels within the window centered around  $p$ .

Horn-Schunk [4] method assumes the flow of pixels across the image in subsequent frames is smooth and distortion free. It approximates the optical flow by solving the equations

$$\begin{aligned} I_x(I_x u + I_y v + I_t) - \alpha^2 \Delta u &= 0 \\ I_y(I_x u + I_y v + I_t) - \alpha^2 \Delta v &= 0 \end{aligned} \quad (8)$$

where  $\Delta = \frac{\delta^2}{\delta x^2} + \frac{\delta^2}{\delta y^2}$  is the Laplacian operator. The above approaches suffer from multiple drawbacks which we discuss below.

- These approaches require calculation or approximation of the higher order polynomials making it computationally inefficient for scaling.
- These approaches are make several assumptions about the images. These assumptions fail to hold in the wild.
- These approaches heavily depend on the constant brightness principle and fail to perform well for images with varying brightness, illumination, distortion and color.

## 5 NEWER APPROACHES

### 5.1 Convolutional Neural Networks

Recent research on continuous motion detection is heavily focused on the use of deep convolutional neural networks for the task object detection. Deep convolutional neural networks [6] are characterized by the movement of the convolution operator over the input image. Each convolution operator is called a filter and the filter is responsible for capturing unique patterns appearing within the input image. Filters use non-linear activation helping the network capture non linear relationships that may exist within the input data. One complete traversal of a filter over the input results in a feature representation. Multiple such representations are stacked at every convolution layer to create feature maps. Feature maps capture the various patterns that occur across images. Further, these feature maps extract patterns at different scales and intensities making the model robust to affine transformations and scale invariance. These feature maps are further mapped to an embedding layer which extracts important features that appear across these feature maps. This embedding is further fed to a fully connected network which is responsible for making output decisions. Similar images or images belonging to the same class have repetitive patterns. These repetitive patterns are captured within the low representation embedding of the image. Thus, neural networks are invariant to scale, illumination and affine transformation.

## 5.2 Object Detection

Object detection is the task of identifying various objects that are present within the image scene. Feature maps capture various object level information at different scales. The approach to object detection is simple. Along with the classification output the network outputs multiple regression outputs, the regression outputs signify the approximated location of an object in a scene. During backpropagation, neurons which generated the feature map for the object are penalized using mean squared error for misidentifying the location of the object. With a large dataset and over multiple iterations, the network learns to identify feature maps corresponding to objects within the scene. As the network is invariant to scale and affine transformations it generalizes well for similar objects with different sizes, shapes and colours under different illumination and environmental constraints.

Multi-object detection is task of identifying multiple objects within the scene. The task is complex as compared to single object detection problem. Traditional approaches involved using a sliding window of fixed dimension across feature maps and feeding each window to the fully-connected network. The network decides whether the window contains an object. This approach has the following drawbacks

- As the window slides per pixel, large number of input vectors are generated and fed to the fully connected network making it computationally inefficient
- Since the size of the window is fixed, it fails to capture objects of varying scale
- Overlapping of windows leads generates large number of false positives.

Girshick et al. [2], use external image processing techniques such as histogram of oriented gradients and pixel wise unsupervised segmentation to generate candidate object boxes, thereby reducing the number of possible input vectors. Ren et al. [10], use the existing annotations available within the dataset to generate the candidate boxes and generate equal random boxes from different part of the image to generate negative examples, classified as background. Redmon et al. [9], propose a single sweep over the input image and train the network for multiple object detection in an end-to-end fashion.

## 5.3 Recurrent Neural Networks

Recurrent neural networks [7] are a variant of traditional neural networks. They are characterized by a recursive loop which feeds the output of the network back as an input to the network. This allows information to persist within the network. Recurrent neural networks have successfully been used for addressing multiple challenges. They are extensively used for natural language modelling, speech recognition, cognitive science and time-series analysis. This is due to their excellent ability to model and memorize sequences for long temporal duration. Vanilla recurrent neural network has two major drawbacks. We discuss them below.

- Recurrent neural networks can have several input units depending upon the temporal duration of the activity. When the input signal flows from one unit within the network to

the next, it is attenuated using non-linear activation. After flowing through several units within the network, the input signal dies off. This problem is traditionally known as the gradient vanishing problem. If the gradient vanishes during the forward pass of the network, the ability of the network to learn via gradient descent through back-propagation is weakened. Thus, the network fails to learn sequences over longer duration.

- Recurrent neural networks also suffer from gradient explosion problems. An unrolled version of a recurrent neural network is equivalent to feed forward neural networks. During backpropagation the gradient generated by the final layer is accumulated over the layers, leading to drastic unstable changes in the weights of the networks.
- Vanilla recurrent neural networks also suffer due to lack of inbuilt correction mechanism. Advanced versions of Recurrent neural networks can correct the weights of the recursive loop to control the influence of previous inputs on the correct input thereby preventing the weights from exploding.

## 5.4 Long Short Term Memory Network

Long short term memory network [3] is an advanced variation of the vanilla recurrent neural network. They overcome the gradient explosion and gradient vanishing drawbacks of the traditional recurrent network. They are characterized by a memory cell which holds sequential information. This characteristic of the network is important in tracking motion of objects in subsequent frames of video. The memory cell is connected to 4 gates which determine the flow of information to and from the memory cell and within the network.

- Forget gate: Forget gate determines the influence previous outputs have on the current unit of the long short term memory network. The decision of the gate is driven by the rest of the network. This is essential, as blocking the flow of information to the memory cell reduces the compounding error, stabilizing the training and preventing the gradients to explode during backpropagation
- Input gate: Input gate determines, the joint influence of the current input and the output of the previous layer has on the unit of the long short term memory network. The decision of the gate is driven by rest of the network. Input gate along with the forget gate determine the sequence the network is tasked to remember.
- Output gate: Output gate determines the influence the current unit has on future units.

## 5.5 Working

The network first determines the importance of the previous output with respect to the state of the current unit. The units within the forget gate  $f_t$  are driven to either 1 or 0 by the rest of the network.

$$f_t = \sigma(W_f[h_{t-1}, x_t]) \quad (9)$$

The network then determines the influence of the input to the current unit of the long short term memory network. The weights of the input gate  $I_t$  unit is driven to either 1 or 0 by the rest of the network.

$$i_t = \sigma(W_i[h_{t-1}, x_t]) \quad (10)$$

$$\hat{C}_t = \tanh(W_c[h_{t-1}, x_t]) \quad (11)$$

$\hat{C}_t$  determines the intermediate update vector of the unit

The memory cell of the current unit is then updated with the new update vector determined by the following equation

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \quad (12)$$

The network outputs the following to drive the future units

$$o_t = \sigma(W_o[h_{t-1}, x_t]) \quad (13)$$

$$h_t = o_t * \tanh(C_t) \quad (14)$$

## 6 APPROACH

We present an end-to-end approach for continuous motion detection and object tracking using deep convolutional neural network and long short term memory networks. As the approach is end-to-end, we can train all networks in a single shot and optimize neural network layer weights using a single loss metric and gradients.

---

### Algorithm 1 Single Shot motion tracking

---

- 1: Load video frames  $f_i$ , ground truth labels  $l_i$ , bounding box annotations  $b_{i,j}$
  - 2: Pre-process  $f_i, l_i, b_{i,j}$
  - 3: **procedure** MODEL( $f_i, l_i, b_i$ )
  - 4: Create arbitrary input batch of 64 frames, labels, annotations
  - 5: Feed batch to CNN
  - 6: Generate embedding  $E_{CNN}$ , annotations  $b_{CNN}$ , label  $l_{CNN}$
  - 7: Concatenate  $E_{CNN}$ ,  $b_{CNN}$ ,  $l_{CNN}$  to form long vector  $V_{CNN}$
  - 8: Feed  $V_{CNN}$  to LSTM
  - 9: Predict approximated location of object in future frame  $b_{+1}$
  - 10: Compute joint loss and gradients  $g$  and update network weights
  - 11: Save Model
- 

Our approach is pretty similar to the one presented by Valipour et al. [11] which was used for end-to-end video segmentation. The approach is simple, we begin by feeding a video frame from an arbitrary position within the video through a deep convolutional network. We concatenate lower dimensional embedding of the input frame along with the class of the object and its location to

form a one dimensional long vector. This vector is then fed to a stacked long short term memory architecture. The output of the long short term memory network is approximated location of the object in the subsequent frame.

[Figure 1 about here.]

We pre-train the deep convolutional neural network for the dataset [12]. Pre-training the neural network helps it to converge faster and stabilizes the training of the long short term memory networks which we introduce further. We believe a pre-trained deep neural network architecture trained on the ImageNet dataset will further improve the performance and generalizability of the model for unseen data samples. We do not present results or analysis for ImageNet trained architectures in this experiment. We use the deep convolutional network to extract a lower level embedding of the input image. Lower level embedding captures features present within an image.

We believe, the class of the object plays an important role in determining the rate of change in motion of the objects in subsequent frames e.g. Divers leaning forward, motion change of a car as compared to bicycles will be higher. Additionally, we use the current location of the object as a correction mechanism for the long short term memory network to prevent it from diverging from the actual motion the object in subsequent frames. We use a time stamp of 5 frames to determine the location of the object in the 6<sup>th</sup> frame. We believe 5 frames captures the relative motion of the object between frames. We jointly try to minimize a multiple loss function.

#### 6.0.1 Cross-entropy loss.

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^n y^{(i)} \ln \delta(x^{(i)}) + (1 - y^{(i)}) \ln (1 - \delta(x^{(i)})) \quad (15)$$

$x_i$  is the input samples and  $y_i$  is the output for the corresponding input samples.  $\delta(x_i)$  is the output of the activation function

$$\delta(x) = \frac{1}{1 + e^{-Wx-b}} \quad (16)$$

$W$  and  $b$  represent the weights and bias of the neural network. We minimize the cross-entropy loss for the classification task.

#### 6.0.2 Mean-square loss.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \delta(x^{(i)}))^2 \quad (17)$$

$x_i$  is the input samples and  $y_i$  is the output for the corresponding input samples.  $\delta(x_i)$  is the output of the activation function

$$\delta(x) = \frac{1}{1 + e^{-Wx-b}} \quad (18)$$

$W$  and  $b$  represent the weights and bias of the neural network. We minimize the mean square error loss for the regression task.

We minimize the cross-entropy and mean squared loss generated by the deep convolutional network in predicting and localizing the object in the video frame. In an end-to-end fashion, we allow the loss generated by the long short term memory network to flow through the convolutional neural network to jointly minimize the loss in predicting the location of the object in the next frames.

## 6.1 Salient Features of Architecture

- The input size of frames vary for different videos. Frame heights are centered on 240 pixels and width on 360 pixels. We use a fixed dimension of 120 x 120 pixels for this experiment. The number of frames per video is high, to prevent computational bottlenecks we use a lightweight convolutional neural network architecture. Another benefit of a light weight architecture is it allows the network to be deployed on low powered devices.
- We use 5 convolutional layers with 32, 64, 128 and 256 filters respectively. Each filter is of the size 3 x 3. We use the Adam optimizer for our experiments.
- with each convolutional pass, we reduce the dimensions of the image by half thereby further improving the computational efficiency of our model. We use strides instead of the traditional pooling architecture. Pooling leads to loss of visual information and deteriorates object localization performance.
- We use Rectified linear units for activation in each of our layers. Rectified units prevent gradient saturation and are computationally efficient to calculate.
- Additionally, as we are dealing with a medium sized dataset, our model is prone to over fitting. To prevent over fitting, we couple every convolutional neural network layer with a dropout layer. Dropout serves as a regularizer for neural network architectures as it randomly drops out multiple neurons in the every hidden layer, leading to high misclassification and mean square error and penalizing the neurons forcing them to generalize for all classes.
- We use affixed batch size of 64 temporal frames for our experiment. We randomly select the starting frame so as to prevent our network from over fitting on the sequence of the video frames.
- We use a 2 layered stacked long short term memory network with 16 and 4 units each. We use hyperbolic tangent as an activation for the recurrent neural network. We use categorical cross-entropy loss for the classification task and we use mean squared error for the regression task

## 6.2 Training

We train our model on Indiana University's cluster computing resource Big Red 2 and Karst. We use a total of 6 CPU's to pre-train our model for the classification and regression task and we use one graphical processing unit along with 6 CPU's and 96 cores to train our end-to-end model. We train our model for 100 iterations. For each iteration during pre-training we randomly sample a batch of 64 images, labels and bounding boxes for the classification and regression task. As we are dealing with a medium sized dataset, the error converges quickly. It takes 6 hours to pre-train our model on the above stated configuration. For end-to-end training we use the following approach. We train the model for 75 iterations. In

Each iteration, we sample a sequence of 128 video frames randomly and pass it to our model to generate future location of our object in context. We repeat this procedure for 100 randomly sampled sequences of 128 video frames. Random sampling improves the generalizability of our approach. Additionally as we sample randomly, we are required to repeat this approach multiple times to ensure our model trains on all action sequence. Training the model in an end-to-end fashion requires 12 hours on CPU configuration and takes about 3 hours with the Graphical Processing Unit.

## 7 EVALUATION

We evaluate multiple metrics during training. We try and minimize the cross-entropy loss for the classification task, and minimize the mean-squared error loss for the object position in current frame and future frame. We achieve a cross-entropy loss of 2 % in relative context. Additionally, each iteration uses a randomly sampled sequence of video frame thus indicating our model generalizes well for different actions being performed in the video. The mean squared error loss for the future frame decreases with every iteration indicating our model is able to learn motion sequence in continuous frame. The loss decreases to a minimum of 5 % in relative context which is significant as the loss is jointly computed for all 4 coordinates of the future frame simultaneously. Figure 1, shows the training loss compared to validation loss. We can clearly infer that the model performs well on the validation set, at times exceeding the training accuracy.

[Figure 2 about here.]

[Figure 3 about here.]

[Table 1 about here.]

[Table 2 about here.]

## 8 CONCLUSION

Object and motion tracking is a difficult task. We presented traditional approaches to motion tracking and their computational flaws. We further presented a low powered light weight approach to tackle object and motion tracking using convolutional neural networks and recurrent neural networks. Our model fails to generalize for certain video sequences and tracks well for others. This we believe is due to reduced training time and and small sized dataset. Use of heavier architecture introduces computation bottlenecks and places high emphasis on object detection. By using recurrent neural networks we compensate for object localization mistakes made by our lightweight convolutional neural network by studying the apparent motion of the object in frames. We present competitive results on difficult benchmark dataset. We emphasize pre trained deeper networks on ImageNet dataset can improve the performance and generalizability of our approach but it loses the essence of the light weight architecture. We propose further improvements can be achieved by combining the best of both worlds. A two stream convolutional neural network with video frames and low resolution computationally efficient optical flow can improve the performance of our approach.

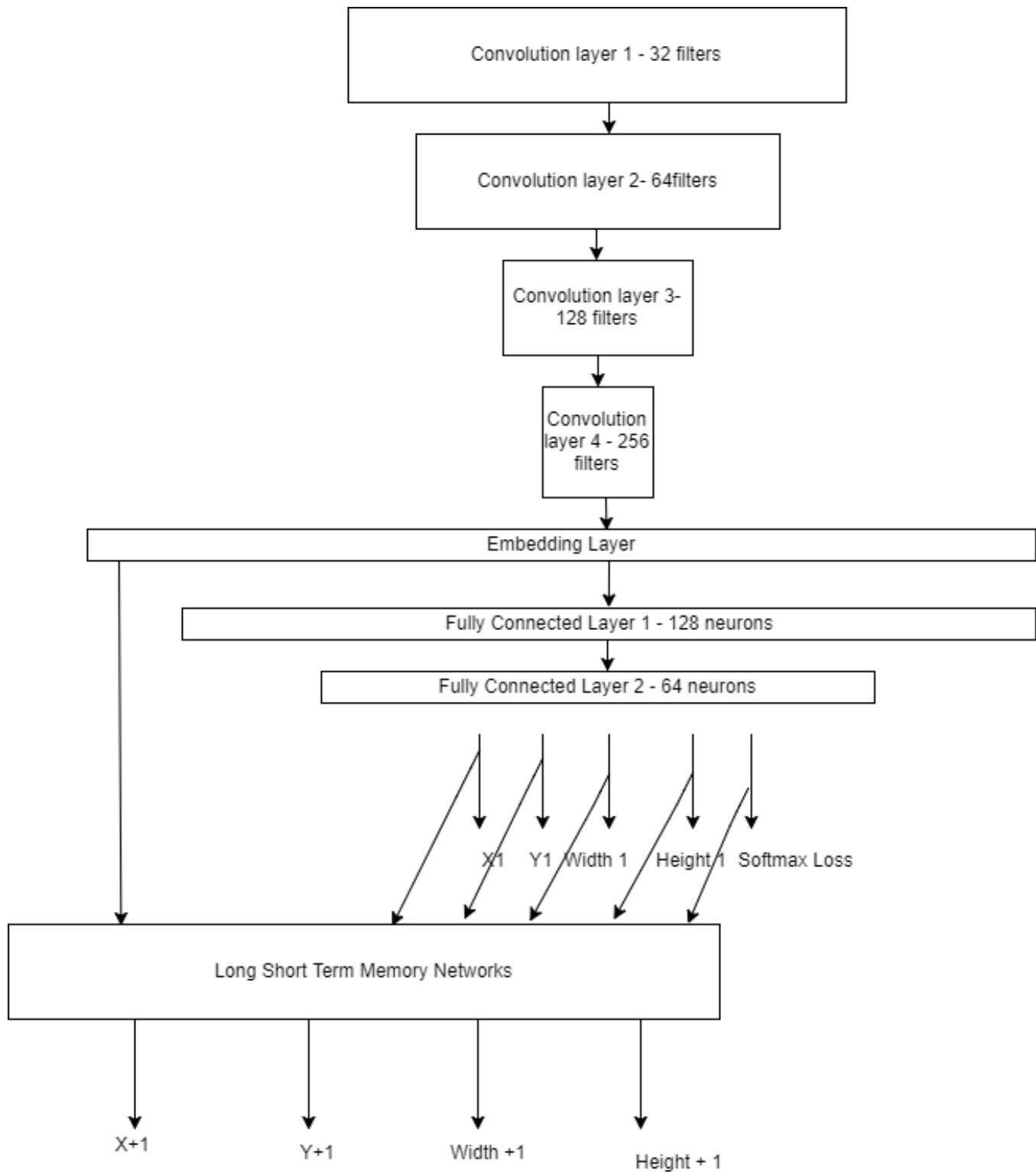
## REFERENCES

- [1] S. Chandana. 2011. Real time video surveillance system using motion detection. In *2011 Annual IEEE India Conference*. 1–6. <https://doi.org/10.1109/INDCON.2011.6107001>.

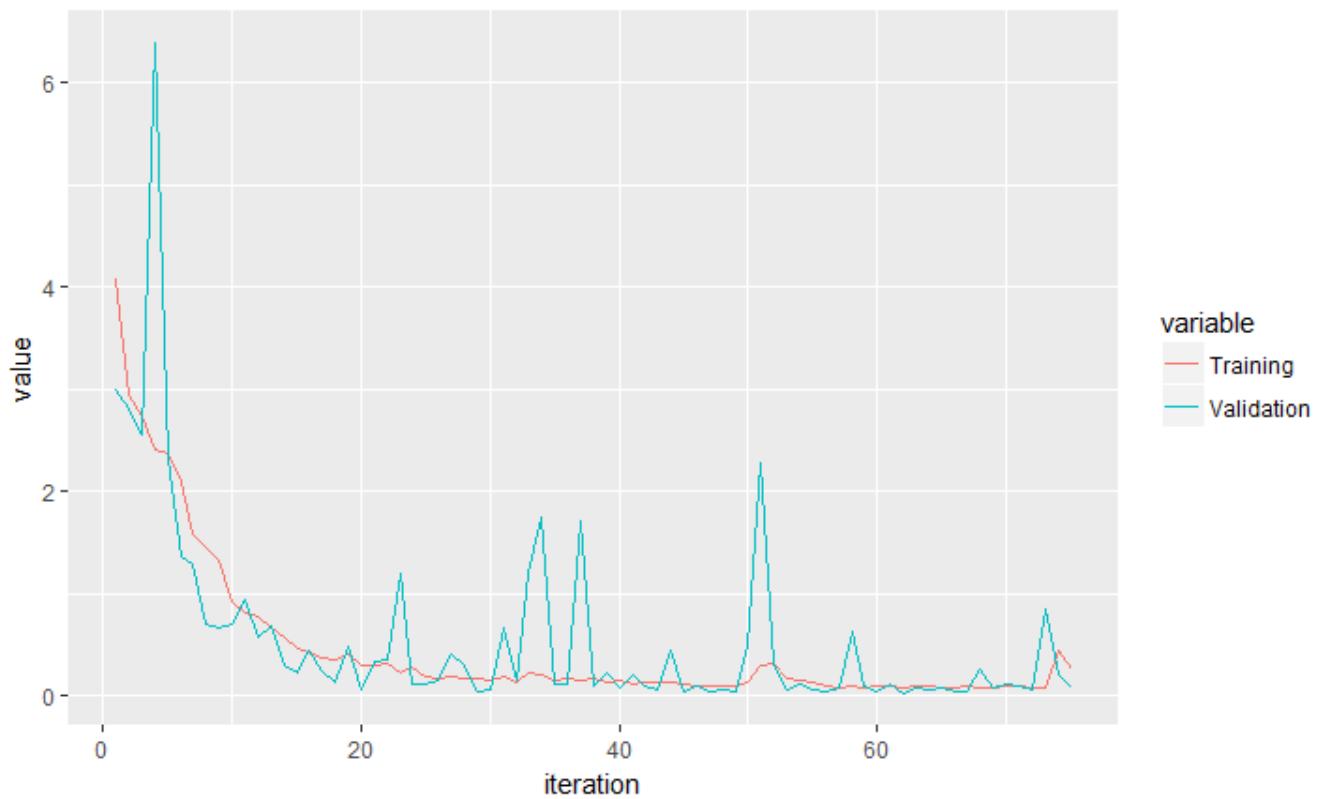
- 2011.6139506
- [2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* abs/1311.2524 (2013). arXiv:1311.2524 <http://arxiv.org/abs/1311.2524>
  - [3] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
  - [4] Berthold K.P. Horn and Brian G. Schunck. 1980. *Determining Optical Flow*. Technical Report. Cambridge, MA, USA.
  - [5] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, T. Vojir, A. Gatt, A. Khajenezhad, A. Salaheddin, A. Soltani-Farani, A. Zarezade, A. Petrosino, A. Milton, B. Bozorgtabar, B. Li, C. S. Chan, C. Heng, D. Ward, D. Kearney, D. Monekosso, H. C. Karaimer, H. R. Rabiee, J. Zhu, J. Gao, J. Xiao, J. Zhang, J. Xing, K. Huang, K. Lebeda, L. Cao, M. E. Maresca, M. K. Lim, M. El Helw, M. Felsberg, P. Remagnino, R. Bowden, R. Goecke, R. Stolk, S. Y. Lim, S. Maher, S. Poullot, S. Wong, S. Satoh, W. Chen, W. Hu, X. Zhang, Y. Li, and Z. Niu. 2013. The Visual Object Tracking VOT2013 Challenge Results. In *2013 IEEE International Conference on Computer Vision Workshops*. 98–111. <https://doi.org/10.1109/ICCVW.2013.20>
  - [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
  - [7] Zachary Chase Lipton. 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. *CoRR* abs/1506.00019 (2015). arXiv:1506.00019 <http://arxiv.org/abs/1506.00019>
  - [8] Janez Pers, Matej Kristan, Matej Perse, and Stanislav Kovacic. 2008. Analysis of Player Motion in Sport Matches. In *Computer Science in Sport - Mission and Methods (Dagstuhl Seminar Proceedings)*, Arnold Baca, Martin Lames, Keith Lyons, Bernhard Nebel, and Josef Wiemeyer (Eds.). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, Dagstuhl, Germany. <http://drops.dagstuhl.de/opus/volltexte/2008/1689>
  - [9] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. *CoRR* abs/1506.02640 (2015). arXiv:1506.02640 <http://arxiv.org/abs/1506.02640>
  - [10] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR* abs/1506.01497 (2015). arXiv:1506.01497 <http://arxiv.org/abs/1506.01497>
  - [11] Sepehr Valipour, Mennatullah Siam, Martin Jägersand, and Nilanjan Ray. 2016. Recurrent Fully Convolutional Networks for Video Segmentation. *CoRR* abs/1606.00487 (2016). arXiv:1606.00487 <http://arxiv.org/abs/1606.00487>
  - [12] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *CoRR* abs/1411.1792 (2014). arXiv:1411.1792 <http://arxiv.org/abs/1411.1792>
  - [13] Jean yves Bouguet. 2000. Pyramidal implementation of the Lucas Kanade feature tracker. *Intel Corporation, Microprocessor Research Labs* (2000).
  - [14] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. 2009. Object tracking using SIFT features and mean shift. 113 (03 2009), 345–352.

#### LIST OF FIGURES

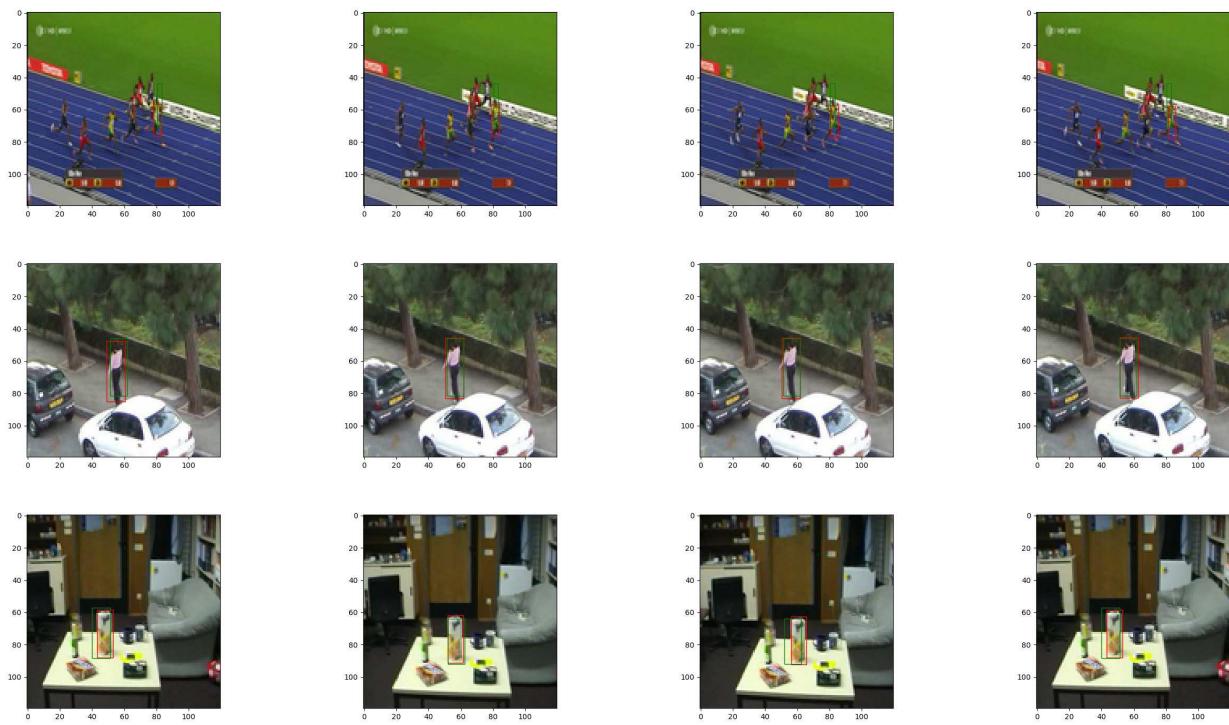
1	Model Architecture	8
2	Train versus validation loss function	9
3	Predicted Examples, green indicates predicted value, red indicates actual location	10



**Figure 1: Model Architecture**



**Figure 2: Train versus validation loss function**



**Figure 3: Predicted Examples, green indicates predicted value, red indicates actual location**

LIST OF TABLES

1	Loss Table-Training	12
2	Loss Table- Validation	12

**Table 1: Loss Table-Training**

Iteration	Training Joint Loss	Training LSTM Loss	Training Cross Entropy Loss
1	2.9984	0.0174	2.7480
10	0.7004	0.0069	0.6378
20	0.0588	0.0035	0.0175
30	0.0470	0.0067	0.0014
40	0.0862	0.0044	0.0372
50	0.4786	0.0120	0.4142
60	0.0349	0.0041	2.3931e-04
70	0.0726	0.0038	0.0141

**Table 2: Loss Table- Validation**

Iteration	Validation Joint Los	Validation LSTM loss	Validation Cross entropy Loss
1	4.0735	0.0296	2.8767
10	0.9193	0.0071	0.8033
20	0.2893	0.0044	0.2122
30	0.1720	0.0035	0.1120
40	0.1459	0.0041	0.0836
50	0.1347	0.0040	0.0678
60	0.1028	0.0034	0.0510
70	0.0745	0.0028	0.0306

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in 6139506
Warning--empty address in 6139506
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Gi
Warning--numpages field, but no articleno or eid field, in Hochreiter:1997:LSM:1246443.1
Warning--empty institution in Horn:1980:DOD:888857
Warning--empty publisher in 6755885
Warning--empty address in 6755885
Warning--empty address in NIPS20124824
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Li
Warning--page numbers missing in both pages and numpages fields in sports
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Re
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Re
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Va
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Yo
Warning--no number and no volume in LK
Warning--page numbers missing in both pages and numpages fields in LK
Warning--no journal in articlezhou
(There were 17 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
report.bib:18:@INPROCEEDINGS{6755885,
report.bib:46:@INPROCEEDINGS{6139506,
```

```
=====
bibtext comma label error
```

```
=====
latex report
```

```
[2017-12-16 09.32.38] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
```

```
bookmark level for unknown defaults to 0.  
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.  
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.  
No positions in optional float specifier. Default added (so using 'tbp').  
No positions in optional float specifier. Default added (so using 'tbp').  
Typesetting of "report.tex" completed in 1.2s.
```

```
./README.yml
```

```
10:81    error    line too long (84 > 80 characters)  (line-length)  
11:81    error    line too long (81 > 80 characters)  (line-length)  
12:81    error    line too long (86 > 80 characters)  (line-length)  
25:81    error    line too long (184 > 80 characters)  (line-length)  
25:183   error    trailing spaces  (trailing-spaces)  
31:28    error    trailing spaces  (trailing-spaces)  
37:81    error    line too long (184 > 80 characters)  (line-length)  
37:183   error    trailing spaces  (trailing-spaces)  
41:14    error    trailing spaces  (trailing-spaces)  
42:81    error    line too long (94 > 80 characters)  (line-length)
```

---

## Compliance Report

---

```
name: Ajinkya Khamkar  
hid: 211  
paper1: 10/27/2017 100%  
paper2: 11/06/2017 100%  
project: Dec 5 2017 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
12  
wc 211 project 12 4434 report.tex  
wc 211 project 12 4614 report.pdf  
wc 211 project 12 943 report.bib
```

```
find "
```

---

```
passed: True
```

find footnote

---

18: \renewcommand\footnotetextcopyrightpermission[1]{} % removes  
footnote with conference information in first column

passed: False

find input{format/i523}

---

passed: False

find input{format/final}

---

passed: False

floats

---

54: Object detection is fundamental and important Computer Vision task. Continuous object detection is an extension of object detection for continuous motion scenes. In section \ref{applications}, we discuss the scope and applications driven by continuous motion detection. In section \ref{experimentaldata}, we present the data that we use for our experiments.

56: In section \ref{traditional}, we discuss traditional techniques used for continuous motion detection. Traditionally, hand crafted features \cite{articlezhou} and optical flow \cite{LK} between subsequent frames was used to detect continuous motion detection. We discuss the major drawbacks of using traditional optical flow and hand crafted feature based methods. These drawbacks can be overcome using newer deep learning techniques.

58: We begin section \ref{newer} by discussing about deep convolutional neural networks and their application for object detection. We introduce certain naive methods that can be used to achieve continuous object detection. In section \ref{RNN}, we introduce Recurrent Neural Networks and their application for continuous object detection. We discuss long-short-term-memory networks a form of recurrent neural network which is designed to retain scene memory over long periods of time. We discuss implementations which use long-short-term-memory networks along with deep convolutional neural networks to improve the performance of the naive methods.

60: In section \ref{approach}, we present an end-to-end approach and  
 algorithm which can be trained in a single shot fashion with the  
 gradient generated by long-short-term-memory network to train the  
 object detection network.  
 62: In section \ref{arch}, we discuss the our training methodology and  
 training resources used for our experiment. In section  
 \ref{results}, we present the results we achieved for training the  
 model in an end-to-end fashion. In section \ref{conclusion}, we  
 conclude our discussion.  
 64: \section{Applications}\label{applications}  
 68: \section{Data} \label{experimentaldata}  
 75: \section{Traditional Approaches} \label{traditional}  
 88: \begin{align} \label{eqn2}  
 92: In equation \ref{eqn2} \$I\$ represents the pixel wise intensity  
 across frames. If the movement is small the right hand side of the  
 above equation can be approximated using the first order Taylor  
 series  
 94: \begin{align} \label{eqn3}  
 98: \begin{align} \label{eqn4}  
 102: \begin{align} \label{eqn5}  
 106: \begin{align} \label{eqn6}  
 110: \begin{align} \label{eqn70}  
 116: \begin{equation} \label{eq7}  
 128: \begin{equation} \label{eq8}  
 147: \section{Newer Approaches} \label{newer}  
 172: \subsection{Recurrent Neural Networks} \label{RNN}  
 204: \begin{align} \label{eqn15}  
 210: \begin{align} \label{eqn16}  
 214: \begin{align} \label{eqn17}  
 222: \begin{align} \label{eqn18}  
 229: \begin{align} \label{eqn19}  
 233: \begin{align} \label{eqn20}  
 238: \section{Approach} \label{approach}  
 244: \caption{Single Shot motion tracking}\label{SSDLSTM}  
 265: \begin{figure}[htbp]  
 266: \includegraphics[width=\linewidth]{images/model.png}  
 268: \label{Model architecture}  
 277: \begin{align} \label{eqn9}  
 283: \begin{align} \label{eqn10}  
 291: \begin{align} \label{eqn11}  
 297: \begin{align} \label{eqn1000}  
 305: \subsection{Salient Features of Architecture} \label{arch}  
 332: \section{Evaluation} \label{results}  
 337: \begin{figure}[htbp]  
 338: \includegraphics[width=\linewidth]{images/Plot.png}  
 340: \label{fig:Evaluation plot}

```
344: \begin{figure}[htbp]
345: \includegraphics[width=\linewidth]{images/Predicted.jpg}
347: \label{fig:Prediction Examples}
351: \begin{table}[]
354: \label{table1}
368: \begin{table}[]
371: \label{table2}
386: \section{Conclusion} \label{conclusion}
```

```
figures 3
tables 2
includegraphics 3
labels 33
refs 6
floats 5
```

```
False : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are referred to: (refs >= labels)
```

Label/ref check

333: We evaluate multiple metrics during training. We try and minimize the cross-entropy loss for the classification task, and minimize the mean-squared error loss for the object position in current frame and future frame. We achieve a cross-entropy loss of 2 \% in relative context. Additionally, each iteration uses a randomly sampled sequence of video frame thus indicating our model generalizes well for different actions being performed in the video. The mean squared error loss for the future frame decreases with every iteration indicating our model is able to learn motion sequence in continuous frame. The loss decreases to a minimum of 5 \% in relative context which is significant as the loss is jointly computed for all 4 coordinates of the future frame simultaneously. Figure 1, shows the training loss compared to validation loss. We can clearly infer that the model performs well on the validation set, at times exceeding the training accuracy.

passed: False -> labels or refs used wrong

When using figures use columnwidth  
[width=1.0\columnwidth]  
do not change the number to a smaller fraction

find textwidth

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in 6139506
Warning--empty address in 6139506
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Gi
Warning--numpages field, but no articleno or eid field, in Hochreiter:1997:LSM:1246443.1
Warning--empty institution in Horn:1980:DOD:888857
Warning--empty publisher in 6755885
Warning--empty address in 6755885
Warning--empty address in NIPS20124824
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Li
Warning--page numbers missing in both pages and numpages fields in sports
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Re
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Re
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Va
Warning--page numbers missing in both pages and numpages fields in DBLP:journals/corr/Yo
Warning--no number and no volume in LK
Warning--page numbers missing in both pages and numpages fields in LK
Warning--no journal in articlezhou
(There were 17 warnings)
```

```
bibtex_empty_fields
```

---

```
entries in general should not be empty in bibtex
```

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
```

---

```
The following tests are optional
```

---

```
=====
```

```
Tip: newlines can often be replaced just by an empty line
```

```
find newline
```

---

```
passed: True
```

```
cites should have a space before \cite{} but not before the {
```

```
find cite {
```

---

```
passed: True
```

# CMD5 Plugin to Create a Docker Swarm Cluster on 3 Raspberry Pis

Andres Castro Benavides  
Indiana University  
107 S. Indiana Avenue  
Bloomington, Indiana 43017-6221  
acastrob@iu.edu

Uma M Kugan  
Indiana University  
107 S. Indiana Avenue  
Bloomington, Indiana 43017-6221  
umakugan@iu.edu

## ABSTRACT

Information technologies are evolving from mainly one-host environments to more distributed environment. Docker Swarm makes it possible to avoid having a single point of failure and instead, have multiple nodes that can be properly balanced and contain replicas of the information. Currently, Dockers must be individually downloaded, installed and configured on each physical computer in order for the desired computers to work in swarm mode. This paper details the development of a plug-in that would allow CloudMesh to deploy a Docker Swarm cluster. The creation of this plug-in would be the first step towards the development of a tool which would allow larger debian based networks to work as container oriented virtual environments with optimized usage of resources.

## KEYWORDS

Raspberry Pi, Cloudmesh, CMD5, Big Data, Big Data, i523, HID305, HID323

## 1 INTRODUCTION

### 1.1 Docker: Swarm mode, Current Use, Installation and Configuration

Docker is the technology used for containerization for software development. It is an open source tool which makes it easy to deploy applications. Applications are packaged in containers and then it is shipped to all the platforms that is supposed to work with. Applications are divided into manageable sizes and all the dependent functions are added and individually packaged. Both Linux and Windows are supported by Docker.

Docker Swarm is a clustering and scheduling tool for Docker containers. A swarm is nothing but multiple Docker hosts which run in swarm mode and act as managers to manage delegation and workers will run swarm services). A given Docker host can be a manager or a worker or it can perform both roles. If any of the worker node becomes unavailable, manager schedules that node's tasks on other nodes. A node is an instance of the docker engine participating in the swarm [5].

A swarm is made up of multiple nodes. We need to execute docker swarm init to enable swarm mode and to make current machine a swarm manager, run docker swarm join on other machines to add them to the swarm as workers and run docker node ls on the manager to view the nodes in this swarm.

Docker Swarms are used to orchestrate processes, optimizing the use of resources across clusters. In other words, the use of Docker Swarms allow individual computers to work as a cluster, sharing their RAM, processors, physical memory, among other features

or abilities. The docker, when used in swarm mode, evaluates the assets across the network and manages tasks in real time. Each computer can contribute its assets to complete tasks in the most efficient way. It is dynamic and adapts based on the available resources and current demands.

In order to set up a Docker Swarm, there needs to be direct access to each machine that will be used as a node (an instance of Docker that will be part of the swarm). In order to set up the nodes, the docker must be independently installed and configured on each machine. Then, each machine must be added to the swarm, allowing it to communicate or interact with the other nodes.

This process not only requires human resources (technicians working on installation and configuration) but also demands these actions be repeated manually on each individual node or manager. While this can be done virtually, it still requires individual attention in the setup of each machine. In order to optimize the setup of Docker Swarms, CloudMesh could be utilized to centralize installation and configuration of every node and manager.

*Inside Docker.* The four main internal components of docker are Docker Client and Server, Docker Images, Docker Registries, and Docker Containers.

*Docker Client and Server.* The docker server gets the request from the docker client and then processes it accordingly. Docker server and docker client can either run on the same machine or a local docker client can be connected with a remote server running on another machine [13].

[Figure 1 about here.]

*Docker Images.* Base images are the operating system images such as Ubuntu 14.04 LTS, or Fedora 20 which creates a container to run the operating system. The docker file contains a list of instructions to build an image. When using docker, we start with a base image, boot up, create changes and those changes are saved in layers forming another image [12].

*Docker Registries.* Docker images are placed in docker registries. It is the same as source code repositories where images can be pushed or pulled from a single source.

*Docker Containers.* Docker image creates a docker container. Containers have everything for the application to run on its own.

#### 1.1.1 Benefits of using Docker.

*Open Source Technology.* The Docker containers are based on open standards which means that anyone can contribute to the Docker tool and at the same time customize it for their needs, if the features they are looking for is not already available.

*Portability.* Docker makes distributed applications to be dynamic and portable which can be run anywhere which makes it extremely popular among developers.

*Sharing.* Docker is integrated with a software sharing and distribution mechanism that allows for sharing and using container content which helps the tasks of both the developer and the operations team.

*Elimination of Environmental Inconsistencies.* Any changes made in one environment will be shared across other environments or all the applications can exist in the single environment.

*Resource Isolation.* Resource isolation adds to the security of running containers on a given host. Docker uses Namespaces technology to isolate work spaces called containers. Namespace is created when container is run and access is limited to that namespace only. Every container in Docker will have its own work space which makes it easier debug if there are issues with any particular container.

*Easy Integration.* Docker can be easily integrated into a variety of infrastructure tools like Amazon Web Services, Ansible, IBM Bluemix, Jenkins, Google Cloud Platform, Oracle Container Cloud Service, Microsoft Azure to name a few.

*Better Security.* Docker provides a interface for developers and IT teams to define and manage their security configurations for applications as it navigates from one stage to another.

*Docker - Use Cases.* The Docker platform is the only container platform to build, secure and manage the variety of applications from development to production both on premises and in the cloud. It also creates room for innovation, increases time to market, highly agile. Docker supports diverse set of applications and infrastructure for both developers and IT. It transforms IT without having to re-tool, re-code or re-vamp existing applications, policies or staff [7].

*DevOps.* The main goal of DevOps is to eliminate the gap between the developers and IT operations team. Docker with DevOps get the developers and operations team to work together so that they both understand the challenges faced by each other, apply DevOps practices [7].

*CI/CD.* Continuous Integration and Continuous Deployment (CI/CD) are the most common use cases of Docker. Continuous Integration testing and Continuous Deployment allows developers to build codes, test them in any environment. Docker integration with Jenkins and GitHub making it easier for developers to build codes, test them in GitHub and trigger a build in Jenkins and adding the image in Docker registries [7].

*Docker Containers As A Service.* Docker help any organization to modernize their application architecture. It can deploy scalable services securely on a wide variety of platforms, improving flexibility and maximizing capacity. Best use case for Docker installation is the US Government where they enhanced their applications and made their components and services of their system and easily transportable/shareable with other agencies within the government [7].

#### 1.1.2 Docker - Services.

*Docker Engine.* Docker Engine is the foundation for the application platform which is used for creating and running Docker containers. It is supported on Linux, Windows, Cloud and Mac OS. It is lightweight, open source and integrated with a work flow to build and containerize applications. User interface is very simple and it makes the environment easily portable from single container on single host to multiple applications on a many number of hosts [7].

*Docker Enterprise.* Docker Enterprise provides an integrated platform for both developers and IT operations team where container management and deployment services are together for end-to-end agile application portability. It is easy to manage, monitor and secure images both within the registry and those deployed across various clusters [7].

*Docker Hub.* Docker Hub functions as a hosted registry service that helps you store, manage, share and integrate images across various developer work flows. Integration testing is done each time when the image is shared [7].

*Docker Compose.* Docker Compose is a tool that developers deploy to define and run all multi-container Docker applications. Single host can be used to isolate multiple environments, even if they are of the same name. Data volume is copied automatically from old container whenever a new container is created. Compose uses the previous configuration to create the new container which reduces the time for replicating the same changes to the environment [7].

## 1.2 CloudMesh

CloudMesh is an innovative tool that allows communication and interaction between cloud based solutions. Not all clouds are docker based and there are different types of virtual and cloud environments. Through CloudMesh, data can be shared and utilized by cloud solutions that are not otherwise programed to communicate with each other. Cloud mesh does not just manage a series of clouds, but centralizes and deploys them as one main system that manages the data resources.

\*\*Quote teacher= Cloudmesh is a project to easily manage virtual machines and bare metal provisioned operating systems in a multicloud environment. We are also providing the ability to deploy platforms\*\*

## 1.3 Creating CloudMesh plug-ins

*what it currently does and has the potential to do.* By creating CloudMesh plug-ins, it is possible to extend its potential from different kinds of cloud based environments interconnection to deployment of a container management system, in this case, Docker .

Utilizing CloudMesh to Centralize Docker Swarm Installation Cloud Mesh does not have a plug in that allows you to deploy container solutions on physical networks. Create a plug in that would allow Cloud Mesh to deploy container solutions (in this case, the Swarm mode of Docker) to a physical Debian based network (in this case, a series of raspberry pies). Could be used as a model to deploy other types of container oriented solutions. It is taking a simple network (Debian based network and allowing it to centralize

resources and assigning tasks and optimizing different functions by installing a container management system, called Docker Swarm.

In order to simulate the deployment of a Docker Swarm cluster, this Cloudmesh project develops a Cloudmesh plug in, that deploys a Docker Swarm cluster on three Raspberry Pi, allowing them to be part of this multi cloud environment.

The cloud mesh allows Methods you to deploy the Docker Swarms (are container management tools) to the raspberry pi's.

## 2 WHAT IS RASPBERRY PI

The Raspberry Pi is a credit-card-sized computer with ARM processor that can run a Linux desktop operating system. Raspberry Pi can plug into TV and a keyboard. It is a little computer which can be used for many of the things that desktop PC does, like spreadsheets, word processing, browsing the internet, playing games and also to play high-definition video. Raspberry is not intended to replace personal computer as its OS support, memory etc are limited when compared to Laptop [10].

### 2.1 Differences between Laptop and a Pi

Raspberry Pi uses an ARM based processor like ARM Cortex A7 or A53 depending upon the model while the traditional PC/Laptop uses a conventional x86 /x64 Processor from either Intel or AMD. Embedded systems had low cost and low power requirements and since ARM processor used in the Raspberry Pi is used in embedded systems, A raspberry pi consumes very much less power than a laptop. The processor is also much slower than most Intel/AMD processors used in PCs, so complex programs can not be executed. Pi does not have any wireless networking capability like WiFi, Bluetooth etc when compared to laptop. Pi comes with 1 GB ram for version 3 while most laptops have 2GB/4GB RAM that can be easily expanded to 16GB. Laptops can have secondary storage for about 1 TB. It also supports Flash based storage which tends to be more expensive per bit than traditional Magnetic Hard drives. Therefore the Raspberry Pi will have a smaller storage capacity than a traditional PC.

## 3 DOCKER ON RASPBERRY PI

Raspberry Pi hardware architecture is called ARM and differs from the architecture behind our regular PC, laptop or cloud instance. A binary built for either system will not execute on the other. Images or binaries that was not created by you or from true source may pose a potential threat. Docker swarm cluster can be built easily on Raspberry Pi with just two basic commands : swarm init and swarm join [6].

## 4 DOCKER AND BIG DATA PLATFORM

It is always been a challenge to maintain or even to have a control deployment environment. It is very difficult to identify any issues without proper deployment environment. Most of the times, issue can be fixed as simple as disabling a service or just uninstalling a software or slightly tweaking the environment. This can be easily achieved only when we have complete control of the environment [9]. It is very difficult to manage a distributed environment whether in cloud or not. There are lot of manual effort whenever there is an installation across multiple nodes. Docker

allows anyone to quickly create, launch and test Docker containers very easily. Container offer lightweight isolation and virtualization, yielding reduced overhead, faster deployments and restarts, and simplified migration. There are lot of frameworks like, Google's Kubernetes, CoreOS, Multi-Container orchestration, etc. which comes in handy with Dockers and Docker is very lightweight when compared to a Virtual Machine. Even though Docker comes very handy in addressing many of these issues, main selling point is building consistent environments which are very easy to replicate. Especially in big data environment, instead of installing every single component from the Hadoop ecosystem, required for their development or testing environment, we can just create it once and use it any number of times and everywhere. Docker allows usage of different versions on the same tool for different jobs without any conflict. Docker containers are a great way of deploying services at scale and giving isolation to services that run on the same host and improving utilization and we can even use Dockers for scheduling batch analytical jobs.

## 5 DOCKER'S PITFALL

Docker was not designed to support the long-running containers that are needed to support production systems. While Docker gets a lot of visibility from the development and DevOps communities, its operational maturity still leaves a big void. There are no logs from containers and hence logging is difficult in a distributed Docker environment. Dockers need separate orchestration, provisioning and automation [3]. Managing a huge amount of containers is challenging, especially when it comes to clustering containers. Running a container need root access and due to security and governance policy, many companies may not grant root privileges for everyone. In some companies, only software from official/trusted sources can be installed on their machines. Since Docker is not included in Red Hat Enterprise Linux 6, it needs to be installed from docker.com, which is an untrusted source [8].

## 6 METHODS: PROPOSED SOLUTION

About the current solution:

This solution was created for a specific type of hardware and software, but is modular enough to be extended to different environments with similar features, such as basic architecture -which include but is not limited to ARM single board computers- and an operating system based on Debian, such as Debian, Raspbian, Ubuntu, etc.

### 6.1 Hardware

For the current proposed solution, the different pieces of hardware were chosen based on criteria such as Compatibility and Price.

The following is a list of the hardware that was used and below that list there is a description of each piece of hardware that was used.

- 3 Raspberry Pi
- 3 Micro SD Cards (64 GB)
- 3 USB to Micro USB Cables for power supply to the Raspberry Pi
- 1 External monitor (for the configuration only).

**6.1.1 Raspberry Pi.** For this experiment, the 3 machines that were used were Raspberry Pi 3 Model B. Raspberry Pi are single boarded computers, that come in a small presentation. They have been developed with education and extension in mind, making them very popular in the academic and entrepreneur communities. The specifications of the model that has been used for this experiment are the following:

- CPU: 1.2 GHZ quad-core ARM Cortex A53 (ARMv8 Instruction Set)
- GPU: Broadcom VideoCore IV @ 400 MHz
- Memory: 1 GB LPDDR2-900 SDRAM
- USB ports: 4
- Network: 10/100 MBPS Ethernet, 802.11n Wireless LAN, Bluetooth 4.0

[2]

The Raspberry Pi are interacting with each other using a private wireless network, and they have been assigned static Internet Protocol Addresses. In this case 192.168.1.85, 192.168.1.86 and 192.168.1.87.

**6.1.2 Micro SD Cards.** Because of its architecture, Raspberry devices require the use of Micro SD Cards to contain the Operative system and other files. They emulate the Hard drive resource used on other kinds of computers. The reason that it is required to have at least 16 GB of memory, is because there will be several pieces of software installed in the devices, each one of them with different requirements:

Docker Memory Requirements [5]:

- 8GB of RAM for manager nodes or nodes running DTR.
- 4GB of RAM for worker nodes.
- 3GB of free disk space.

So at least 12 of the GB would be required for Docker and 4 GB used for the proper functioning of Raspbian. [11]

Taking these requirements in consideration, there should be a minimum of 16GB of free space in the MicroSD in order to perform this experiment.

The Micro SD cards used were San Disc Memory Cards with a 64GB capacity.

**6.1.3 Micro USB Cables.** 3 USB to Micro USB Cables for power supply to the Raspberry Pi Since these small computers don't use the regular power supply chords, they are equipped with MicroUSB ports to power the device. All of these devices are plugged to a main power outlet that allows to charge multiple devices at the same time. There are other options to power the devices include, such as attaching them to external batteries.

**6.1.4 External monitor.** Since the Raspberry Pi are headless machines, they require to be accessed directly for the initial set up and after that it is possible to continue the configuration and installation process using any kind of remote access, like SSH or RealVNC. For this initial connection, any kind of screen that is HDMI compatible is useful. In this case the initial setup of the Raspberry Pi was performed on a Toshiba 55 inch HDTV with HDMI port. After that they were accessed from a Laptop computer with Linux Ubuntu 17.10, using Remmina via ssh (XORG).

**6.1.5 Initial input devices.** In order to set up the devices. The Raspberry Pi will require a set of initial input devices attached to each computer. For this exercise, a USB enabled standard keyboard and a USB enabled standard mouse were used.

## 6.2 Software

**6.2.1 Raspbian.** Currently, the default way to deploy the operating system to the Raspberry Pi is by using an Operating System installation Manager called Noobs -which stands for fiNew Out Of Box Softwarefi-. This manager can be downloaded directly from the Raspberry Pi website and it includes several Operating system options, among them:

- Raspbian
- Pidora
- LibreELEC
- OSMC
- RISC OS
- Arch Linux

Since Raspbian is the default Operating system and most commonly used, this experiment decided to use it. This is also helpful because there is material available in different websites with instructions on how to install Docker in Debian based Machines. Raspbian is Debian based. Another important reason is that Docker has as a requirement that the Linux kernel version on which it will be installed is 3.10 or higher. The Kernel version of the version of Raspbian that was used is 4.9.

The version of Raspbian that was used has the following specifications:

- **Kernel version:** 4.9
- **Release date:** 2017-11-16

**6.2.2 Docker.** There are several versions of Docker available. Each version with their own advantages and disadvantages. Because of the architecture used by Raspberry Pi -ARM instead of AMD-, the Docker version used is **Docker for Debian ARM**. With the following Specifications:

- Version 17.09.0-ce
- Release 2017-09-26

This version of Docker is Community Edition (CE), which means that it is available for free and can be installed on bare metal or cloud infrastructure. This flexibility is good for the experiment, because it will be installed on Raspberry Pi, which are considered physical devices or bare metal Machines. [5]

## 7 PREREQUISITES

There are several reasons to have the pre requisites that the user will find in this document. They will be explained in a separate section. Before using the proposed solution, the userfis environment needs to meet the following requirements:

**7.0.1 Raspbian Installed.** Raspbian must be installed and configured on all Micro SD Cards. For this, the user may download Noobs from <https://www.raspberrypi.org/> and copy it to a formatted Micro SD Card. Once the Raspberry Pi has the MicroSD loaded with noobs in place and has the input devices and display

attached to it, the user may follow the OS installation guide found on: <http://raspbian.org/>

It is advisable to be hooked up to the network where the user is planning on implement this solution before running Noobs for the first time. This will allow the user to download newer packages or Raspbian and avoid interruptions in the process.

This requirement exists because there is a function that is being explored to capture Raspberry Pi images to be deployed later on and avoid the present pre requisite, but it is not ready yet.

**7.0.2 Update OS repositories.** In order to ensure that the user is accessing the latest version available of the software, it is important to update the Raspbian repositories. In this case, the user can access the Terminal and enter the following commands:

"**sudo apt-get update**" to update the list of available repositories and then "**sudo apt-get upgrade**" to upgrade the available packages.

The first time that the user runs one of these commands, the root password will have to be entered. This process might take a few minutes. [4]

**7.0.3 Remote access setup.** Enable SSH on the Raspberry Pi. After Raspbian installation, enable SSH on all your Raspberry Pi machines.

To do this, the user has to add a line in the file "**sshd\_config**" found in the directory "**/etc/ssh/**" The line has to go at the end of in the "**Authentication section**". It has to contain the following string "**PermitRootLogin yes**". [1]

**7.0.4 Changing hostnames.** In order to keep the three Raspberry Pi organized it is highly advisable to assign an exclusive and distinctive hostname to each Raspberry Pi. The three Raspberry Pi have the following static IP addresses:

- (1) pi85 - 192.168.1.85
- (2) pi86 - 192.168.1.86
- (3) pi87 - 192.168.1.87

By default, all Raspberry Pi devices will have the same Host Name.

To change this feature on each machine, the user will have to modify the line that contains "**127.0.1.1**" and as hostname it includes the string "**raspberrypi**" in "**/etc/hosts**" file, in most of the cases it is the last line in the file. Then, the user may type the desired hostname instead of the word raspberrypi and save the file and close it. This part can be done by using the text editor that comes by default with Raspbian, an editor called "**nano**". It is not advisable for the users to modify the rest of the entries, at least as part of this project.

Once the file is modified, the user will have to initialize the hostname with the `hostname.shf` script this can be done using the following line in the Terminal: "**sudo /etc/init.d/hostname.sh**"

To check if the modification has worked as expected, the user may check the hostname of the machine from the Terminal by running the command: "**hostname -I**"

## 8 STEPS FOLLOWED

### 8.1 Testing shell commands prior to integrations with Cloudmesh

Since Raspberry pi is not currently listed under the supported operative systems for Docker or Cloudmesh, The process of deploying Docker and configuring the swarm Mode was successfully tested on the Raspberry Pi first using the commands that are intended for Debian. Once the Swarm was configured, the three Raspberry Pi devices were left on for over 24 hours and it was not observed any kind of abnormal behavior, like looping services in the OS or overheating.

### 8.2 Purchasing the hardware

The different hardware components were purchased via Amazon.com and took anywhere between 2 to 5 days to arrive. The different components can also be purchased through multiple on line sources or local electronics stores.

### 8.3 Installing the components via ssh into every node.

The following steps were followed on each device: Usig the TV as an external monitor, the USB input devices (keyboard and mouse), and the Raspberry Pi with Raspbian installed. An ssh key was generated and the device was accessed using Remmina via a XORG connection from a computer equipped with Linux Ubuntu 17.10 (Artful Aardvark). The components were installed in the following order:

Updated the **Raspbian** packages Installed **Python 3.6.2** and **Python 2.7.13** via PIP and also Installed **Cloudmesh**: following the instructions found in: <https://github.com/cloudmesh/> Installed **Docker CE ARM** via Terminal using the command `curl -sSL https://get.docker.com -sh` as suggested in <https://www.raspberrypi.org/>

### 8.4 Installing and configuring Docker Swarm

**8.4.1 Manager.** Since Docker requires at least one computer to be a Manager and Cloudmesh also requires at least one main configured piece of equipment, a Raspberry Pi was chosen to be the main device (in this case, the Raspberry Pi with the IP address 192.168.1.85). The following command was run on the Terminal or that device to set it as the manager: **sudo docker swarm init --advertise-addr 192.168.1.85**

### 8.5 Workers

The other two Raspberry Pi devices (in this case, the Raspberry Pi with the IP address 192.168.1.86 and the one with 192.168.1.87) were defined as simple worker nodes. To define the workers, the following command was used: **sudo usermod -a -G docker USER** and to work as part of the swarm the command used was: **docker swarm join --token \*\*\* 198.168.1.85:2377** As a last step, it was confirmed that all the nodes were added by using the following command: **sudo docker node ls**

## 8.6 Additional Research

**8.6.1 Other functions considered.** Initially, for this case, it was considered an option to developed a function called CaptureImage and a second function called DeployRaspbian. As their names suggest, the first one intended to capture an image or backup of a Raspberry Pi. This first function would receive the IP address or hostname of the desired machine and the desired location to store the captured image, alongside the corresponding credentials and wrap a `dd` shell command similar to the following:

**`dd if=/dev/mmcblk0 bs=1M -> dd of=imageDir`**

Among the challenges faced, this line was returning an invalid syntax, most likely because of the use of the variables. Since there was not a lot of time, the team decided to postpone this function.

The second function was called DeployRaspbian and would receive the route and name where the image would be deployed (I.e. `/dev/bkp`) and image name and route (I.e. `/Desktop/raspbian.gz`). The shell command that would be wrapped would be:

**`gzip -dc diskNm - sudo dd of=imageName bs=1m conv=noerror,sync`**

More information on this topic can be found in the section called **Backup [www.raspberrypi.org](http://www.raspberrypi.org)**.

Among the challenges, there is no clarity on whether the image can be deployed over a lan connection and this point there is not enough time to run tests. Also, since this is a copy of a previously used Raspbian, there is a chance that there might be conflicts related to the IP addresses that might be stored in different files of the OS.

### 8.6.2 Final code.

## 9 CONCLUSIONS

1. It is possible to create the plug in. 2. Since this was 3. The fact that the passwords would have to be either hard coded or transferred in plain text has to be seen as a vulnerability, that has to be addressed either by adding an encryption/decryption module or finding another way to safely access the root of the target device.

### 9.1 Evaluation

## 10 OTHER OPTIONS CONSIDERED

Other options of coding were considered during the development of this solution. Since all of the deployment can successfully be done via terminal in Raspbian, two main options were considered:

Option 1. A bash script for every part of the deployment and wrap it in python. This option would have been less dynamic and wouldn't make the best use of the available resources, but at the same time it could have been easier to adapt to linux Operating systems other than Raspbian. Option 2. Use the SH subprocess included in python 2.5-3.5. This is the option that was chosen by the team.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions on this project.

## REFERENCES

- [1] Laura Bailey, Laura Novich, Tim Hildred, and David Jorm. 2012. Red Hat Customer Portal. (2012). [https://access.redhat.com/documentation/en-us/red-hat\\_enterprise\\_linux/6/html/v2v-guide/preparation.before.the\\_p2v\\_migration-enable\\_root\\_login\\_over\\_ssh](https://access.redhat.com/documentation/en-us/red-hat_enterprise_linux/6/html/v2v-guide/preparation.before.the_p2v_migration-enable_root_login_over_ssh)
- [2] Brian Benchoff. 2016. Introducing the Raspberry Pi 3. (Feb 2016). <https://hackaday.com/2016/02/28/introducing-the-raspberry-pi-3/>
- [3] Sam Charrington. 2015. Running Hadoop on Docker, in Production and at Scale. (March 2015). <https://thenewstack.io/running-hadoop-docker-production-scale/>
- [4] Debian. 2017. DebianPackageManagement. (2017). <https://wiki.debian.org/DebianPackageManagement>
- [5] Docker. 2017. Docker CE release notes. (Dec 2017). <https://docs.docker.com/release-notes/docker-ce/>
- [6] Alex Ellis. 2016. 5 things about Docker on Raspberry Pi. (Sep 2016). <https://blog.alexisellis.io/5-things-docker-rpi/>
- [7] Hackeroon. 2017. Docker-the Popular Containerization Technology for an Effective Software Development. (2017). <https://hackeroon.com/docker-the-popular-containerization-technology-for-an-effective-software-development-4e2ddc5a329>
- [8] Philipp Hauer. 2015. Discussing Docker. Pros and Cons. (Oct 2015). <https://blog.philipp-hauer.de/discussing-docker-pros-and-cons/>
- [9] Vivek Murugesan. 2015. Why we chose Docker to build our data processing platform. (August 2015). <http://bigdata-madesimple.com/why-we-chose-docker-to-build-our-data-processing-platform/>
- [10] Raspberry Pi and Raspberry Pi-Teach. [n. d.]. FAQS. ([n. d.]). <https://www.raspberrypi.org/help/faqs/>
- [11] Raspberry Pi and Raspberry Pi-Teach. [n. d.]. SD cards. ([n. d.]). <https://www.raspberrypi.org/documentation/installation/sd-cards.md>
- [12] Babak Bashari Rad, Harrison John Bhatti, and Mohammad Ahmadi. 2017. An Introduction to Docker and Analysis of its Performance. *International Journal of Computer Science and Network Security (IJCSNS)* 17, 3 (March 2017), 228. [http://paper.ijcsns.org/07\\_book/201703/20170327.pdf](http://paper.ijcsns.org/07_book/201703/20170327.pdf)
- [13] James Turnbull. 2014. *The Docker Book: Containerization is the new virtualization*. James Turnbull, New York, USA.

## A WORK BREAKDOWN

**Introduction-Docker, Docker and Big Data Platform, Docker's Pitfall, Raspberry Pi, Difference between Pi and Laptop, Dockers on Pi**  
Uma Kugan.

**Introduction - Cloudmesh** Andres Castro Benavides.

**Editing:** Andres Castro Benavides and Uma Kugan.

LIST OF FIGURES

1 Docker Architecture [13]

9

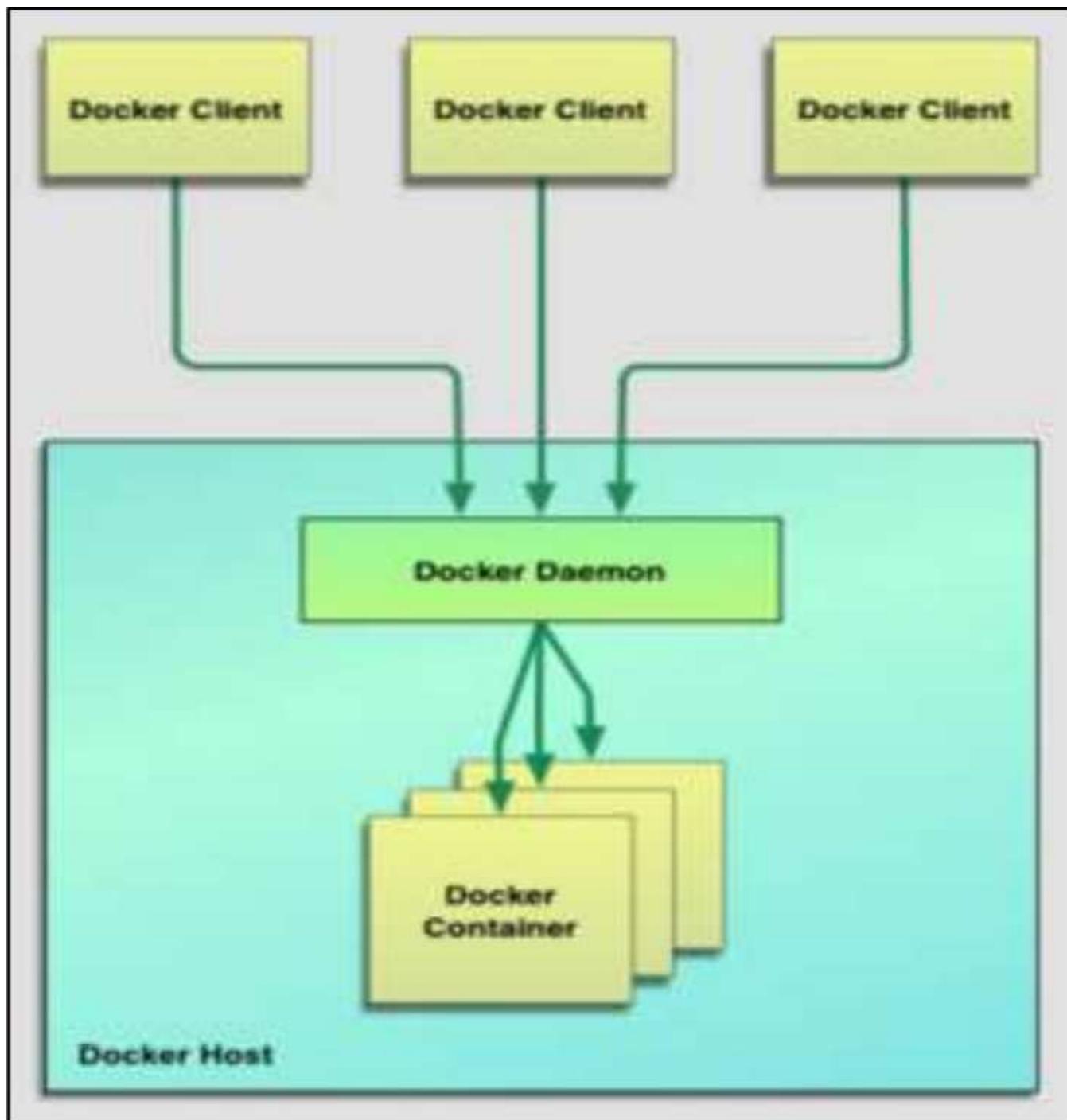


Figure 1: Docker Architecture [13]

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty year in Rpi
Warning--empty year in rpicards2017
(There were 2 warnings)
```

```
bibtext _ label error
```

```
=====
bibtext space label error
```

```
=====
report.bib:127:@misc{Rpi,
report.bib:32:@misc{rpicards2017,
report.bib:40:@misc{dockerdoc2017,
report.bib:50:@misc{noobs,
report.bib:59:@misc{Raspbian,
report.bib:68:@misc{debianpackage,
report.bib:77:@misc{rootSsh,
report.bib:85:@misc{hackaday2016,
```

```
bibtext comma label error
```

```
=====
latex report
```

```
=====
[2017-12-16 09.37.11] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
```

```
Missing character: ""
Missing character: ""
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.1s.
./README.yml
20:81    error    line too long (85 > 80 characters)  (line-length)
24:1     error    trailing spaces  (trailing-spaces)
54:1     error    trailing spaces  (trailing-spaces)
61:12    error    too many spaces after colon  (colons)
61:81    error    line too long (81 > 80 characters)  (line-length)
63:81    error    line too long (83 > 80 characters)  (line-length)
63:83    error    trailing spaces  (trailing-spaces)
64:81    error    line too long (87 > 80 characters)  (line-length)
64:87    error    trailing spaces  (trailing-spaces)
65:81    error    line too long (86 > 80 characters)  (line-length)
65:86    error    trailing spaces  (trailing-spaces)
66:81    error    line too long (85 > 80 characters)  (line-length)
66:85    error    trailing spaces  (trailing-spaces)
67:81    error    line too long (87 > 80 characters)  (line-length)
67:87    error    trailing spaces  (trailing-spaces)
68:81    error    line too long (86 > 80 characters)  (line-length)
68:86    error    trailing spaces  (trailing-spaces)
69:81    error    line too long (95 > 80 characters)  (line-length)
69:94    error    trailing spaces  (trailing-spaces)
70:81    error    line too long (87 > 80 characters)  (line-length)
71:81    error    line too long (92 > 80 characters)  (line-length)
71:92    error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Uma M Kugan
hid: 323
paper1: Review Date 11.10.2017
paper2: Review Date 11.06.2017
project: Dec 14 17 80%
```

yamlcheck

---

wordcount

---

9

```
wc 323 project 9 4957 report.tex
wc 323 project 9 4945 report.pdf
wc 323 project 9 281 report.bib
```

find "

---

```
345: \textbf{\textit{dd if=/dev/mmcblk0 bs=1M | gzip -" | dd
      of=imageDir}}\\
```

passed: False

find footnote

---

passed: True

find input{format/i523}

---

```
6: \input{format/i523}
```

passed: True

find input{format/final}

---

passed: False

floats

---

```
63: \begin{figure}
65: \includegraphics[width=1.0\columnwidth]{images/docker_architecture
    .png}
66: \caption{Docker Architecture} \cite{turnbull2014docker}
    \label{fig:figure1}
```

figures 1

tables 0

includegraphics 1

labels 1

```
refs 0
floats 1
```

```
True : ref check passed: (refs >= figures + tables)
True : label check passed: (refs >= figures + tables)
True : include graphics passed: (figures >= includegraphics)
False : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
passed: True
```

```
When using figures use columnwidth
[width=1.0\columnwidth]
do not change the number to a smaller fraction
```

---

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty year in Rpi
Warning--empty year in rpicards2017
(There were 2 warnings)
```

---

```
bibtex_empty_fields
```

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

non ascii found 8217  
non ascii found 8220  
non ascii found 8221  
non ascii found 8217  
non ascii found 8220  
non ascii found 8221

=====

The following tests are optional

=====

Tip: newlines can often be replaced just by an empty line

find newline

---

passed: True

cites should have a space before \cite{} but not before the {

find cite {

---

passed: True