

# *Use Cases in Big Data Software and Analytics*

Vol. 1, Fall 2017

---

*Bloomington, Indiana*

Sunday 10<sup>th</sup> December, 2017, 14:29

Editor:  
Gregor von Laszewski  
Department of Intelligent Systems  
Engineering  
Indiana University  
laszewski@gmail.com

# Contents

<b>1 Preface</b>	<b>7</b>
1.1 Disclaimer . . . . .	7
1.2 Citation . . . . .	7
1.3 List of Papers . . . . .	8
<b>2 Biology</b>	<b>11</b>
<b>3 Business</b>	<b>11</b>
4 hid202	Status: Dec 04 17 100%
Big Data Analysis in E-Commerce	
Himani Bhatt, Mrunal Chaudhary . . . . .	11
5 hid233	Status: 10%
Big Data in Safe Driver Prediction	
Wang, Jiaan, Chaturvedi, Dhawal . . . . .	46
6 hid234	Status: 100% Dec 07 17
Big Data Analytics and Applications in the Travel Industry and its Potential in Improving Travel Accessibility	
Weixuan Wang . . . . .	60
7 hid235	Status: unkown
Big Data analytics in predict house price	
Yujie Wu . . . . .	76
8 hid301	Status: 100% 12/2/2017
Importance of Big data in predicting stock price	
Gagan Arora . . . . .	88
9 hid306	Status: 100%; 12/3/2017
Predicting Housing Prices - Kaggle Competition	
Murali Cheruvu, Anand Sriramulu . . . . .	100
10 hid320	Status: 100% Dec 03 2017
Big Data Applications in Real Estate Analysis	
Elena Kirzhner . . . . .	139
11 hid324	Status: Dec 04 17 100%
Big Data Analytics in factors affecting Bitcoin	
Ashok Kuppuraj . . . . .	165
12 hid328	Status: Dec 4 17 100%
Predicting Profitable Customers in Banking Industry	
Dhanya Mathew . . . . .	187

13 hid329	Status: 100% Dec 4	
Big Data and The Customer Experience Journey		
Ashley Miller . . . . .	208	
<b>4 Edge Computing</b>	<b>230</b>	
14 hid201	Status: 100%	
IoT Application Using MQTT and Raspberry Pi Robot Car		
Arnav, Arnav . . . . .	230	
15 hid316	Status: 100%	
Big Data and Edge Analytics in Weather Monitoring and Forecasting		
Robert Gasiewicz . . . . .	243	
16 hid319	Status: 80%	
Face Detection and Recognition Using Raspberry Pi Robot Car		
Mani Kumar Kagita . . . . .	244	
17 hid334	Status: Dec 04 17 100%	
The Intersection of Big Data and IoT		
Peter Russell . . . . .	262	
<b>5 Education</b>	<b>274</b>	
18 hid236	Status: Dec 4 17 - 100%	
Big Data and Its Application in Education		
Weipeng Yang, Geng Niu . . . . .	274	
<b>6 Energy</b>	<b>296</b>	
<b>7 Environment</b>	<b>296</b>	
19 hid330	Status: 100%	
Big Data Analytics in Monitoring Outdoor Air Quality		
Janaki Mudvari Khatiwada . . . . .	296	
20 hid345	Status: 100%	
Agricultural Data Science		
Ross Wood . . . . .	323	
21 hid346	Status: Dec 08 17 100%	
Big Data Analysis for Wild File Prevention and Tracking		
Zachary Meier . . . . .	342	
<b>8 Government</b>	<b>349</b>	
22 hid310	Status: Dec 04 17 100%	
Gerrymandering Detection Using Data Analysis		
Kevin Duffy . . . . .	349	
<b>9 Health</b>	<b>374</b>	
23 hid232	Status: 0%	
Big Data and Hearing Disabilities		
Rahul Velayutham . . . . .	374	

24 hid237	Status: 100%, Dec 7, 2017	
	Analyzing everyday challenges of people with visual impairments	
	Tousif Ahmed . . . . .	388
25 hid311	Status: 100%	
	Big Data in Genomics and Medicine	
	Matthew Durbin . . . . .	416
26 hid313	Status: 100%	
	The Impact of Clinical Trial Results on Pharmaceutical Stock Performance	
	Tiffany Fabianac . . . . .	430
27 hid327	Status: 100% 12/05/17	
	How Big Data Will Help Improve People's Health Worldwide	
	Paul Marks . . . . .	451
28 hid331	Status: Dec 4 17 100%	
	Big Data Applications in Predicting Hospital Readmissions	
	Tyler Peterson . . . . .	467
29 hid332	Status: 100%	
	Big Data Analytics to Reduce Health Care in the United States	
	Judy Phillips . . . . .	484
30 hid335	Status: 100%	
	Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics	
	Sean Shiverick . . . . .	499
31 hid337	Status: Dec 04 17 100%	
	IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)	
	Ashok Reddy Singam, Anil Ravi . . . . .	531
32 hid339	Status: Dec 2 2017 100%	
	Diagnosis of Coronary Artery Disease Using Big Data Analysis	
	Hady Sylla . . . . .	550
33 hid348	Status: 100%	
	Big Data Application in Precision Medicine and Pharmacogenomicsn	
	Budhaditya Roy . . . . .	563
<b>10 Lifestyle</b>		<b>586</b>
34 hid109	Status: 100% Dec 4th	
	Diversification of Big Data	
	Shiqi Shen, Qiaoyi Liu . . . . .	586
35 hid231	Status: 100% Dec 4, 2017	
	Big Data Analytics on Food Products Around the World	
	Vegi, Karthik, Chandwani, Nisha . . . . .	611
36 hid302	Status: 100%	
	Recipe Ingredients Analysis	
	Sushant Athaley . . . . .	637
<b>11 Machine Learning</b>		<b>678</b>

37 hid209	Status: Dec 04 17 100%
Comparison between different classification algorithms in Digit Recognizer	
Han, Wenxuan, Liu, Yuchen, Lu, Junjie . . . . .	678
38 hid343	Status: 100 %
Income Prediction Using Machine Learning Techniques	
Borga Edionse Usifo . . . . .	699
<b>12 Media</b>	<b>726</b>
39 hid230	Status: unkown
Big data with natural language processing	
Yuanming Huang . . . . .	726
40 hid340	Status: Dec 7 17 66%
New Approaches to Managing Metadata at Scale in Research Libraries	
Timothy A. Thompson . . . . .	726
<b>13 Physics</b>	<b>731</b>
41 hid304	Status: Dec 04 17 100%
How Far have Space Walks Walked	
Ricky Alan Carmickle . . . . .	731
<b>14 Security</b>	<b>731</b>
42 hid224	Status: Dec 04 17 100%
Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks	
Rawat, Neha . . . . .	731
<b>15 Sports</b>	<b>748</b>
43 hid228	Status: Dec 04 17 100%
Big data applications in Indian Premier League	
Swargam, Prashanth . . . . .	748
<b>16 Technology</b>	<b>765</b>
44 hid308	Status: 0%
TBD	
Pravin Deshmukh . . . . .	765
<b>17 Text</b>	<b>765</b>
<b>18 Theory</b>	<b>765</b>
<b>19 Transportation</b>	<b>765</b>
<b>20 TBD</b>	<b>765</b>
45 hid101	Status: Dec 09 17 100%
TBD	
Huiyi Chen, Yuanming Huang . . . . .	765

46	hid102	TBD	Dianprakasa, Arif	Status: Dec 04 17 0%	778
47	hid107	Big Data Analytics in Support Filtering Wrong Informations On Social Networking Sites	Ni,Juan	Status: Dec 08 0600 100%	778
48	hid111	TBD	Lewis, Derek	Status: unknown	800
49	hid314	TBD	Sarang Fadnavis	Status: Dec 04 17 0%	800
50	hid318	None	Irey, Ryan	Status: Dec 04 17 0%	800
51	hid321	TBD	Knapp, William	Status: Dec 04 17 0%	800
52	hid323	None	Uma M Kugan	Status: Dec 04 17 0%	800
53	hid326	None	Mohan Mahendrakar	Status: unknown	810
54	hid336	None	Jordan Simmons	Status: Dec 04 17 0%	810
55	hid341	Not submitted	Tibenkana, Jacob	Status: 0%	810
56	hid342	TBD	Nsikan Udojen	Status: 0%	810

# Chapter 1

## Preface

### 1.1 Disclaimer

The papers provided are contributed by students of the i523 class thought at Indiana University in Fall of 2017. The students were educated in plagiarizm and we hope that all papers meet the high standrads provided by the policies set at Indiana University in regrads to plagiarizm. In case you notice any issues, please contact Gregor von Laszewski (laszewski@gmail.com) so we cn address the issue with the student.

### 1.2 Citation

The proceedings is at this time available as a draft. To cite this proceedings you can use the following citation entry:

```
@Book{las17-i523,
  editor = {Gregor von Laszewski},
  title = {Use Cases in Big Data Software and Analytics},
  publisher = {Indiana University},
  year = {2017},
  volume = {1},
  series = {i523},
  address = {Bloomington, IN},
  edition = {1},
  month = dec,
  url={https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-i
}
```

Contributors to the volume can cite their contribution as follows. They just need to *FILLIN* the missing information

```
@InBook{las17-,
```

```

author =      {FILLIN},
editor =      {Gregor von Laszewski},
title =       {Use Cases in Big Data Software and Analytics},
chapter =     {FILLIN},
publisher =   {Indiana University},
year =        {2017},
volume =      {1},
series =      {i523},
address =     {Bloomington, IN},
edition =     {1},
month =       dec,
url={https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-i
pages =       {FILLIN},
}

```

## 1.3 List of Papers

HID	Author	Title
101, 230	Huiyi Chen, Yuanming Huang	TBD
102	Dianprakasa, Arif	TBD
105	Lipe-Melton, Josh	Predictive Model For English Premier League Games
107	Ni,Juan	Big Data Analytics in Support Filtering Wrong Informations On Social Networking Sites
109, 106	Shiqi Shen, Qiaoyi Liu	Diversification of Big Data
111	Lewis, Derek	TBD
201	Arnav, Arnav	IoT Application Using MQTT and Raspberry Pi Robot Car
202, 205	Himani Bhatt, Mrunal Chaudhary	Big Data Analysis in E-Commerce
209, 213, 214	Han, Wenxuan, Liu, Yuchen, Lu, Junjie	Comparison between different classification algorithms in Digit Recognizer
211	Khamkar, Ajinkya	Continuous motion tracking using Deep Neural Networks and Recurrent Neural Networks
212, 225, 210	Kumar, Saurabh; Schwartzer, Matthew; Hotz, Nicholas	Can Blockchain Adoption Mitigate the Opioid Crisis Through More Secure Drug Distribution?
215, 208	Mallala, Bharat, Jyothi Pranavi Devineni	Big Data Analytics on Influencers in Social Networks
219	Syam Sundar Herle	Unsupervised Learning for detecting fake online reviews
224	Rawat, Neha	Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks
228	Swargam, Prashanth	Big data applications in Indian Premier League
229	ZhiCheng Zhu	Big Data Analytics in Product Development Management
230	Yuanming Huang	Big data with natural language processing

231, 203	Vegi, Karthik, Nisha	Chandwani,	Big Data Analytics on Food Products Around the World
232	Rahul Velayutham		Big Data and Hearing Disabilities
233, 204	Wang, Jiaan, Dhawal	Chaturvedi,	Big Data in Safe Driver Prediction
234	Weixuan Wang		Big Data Analytics and Applications in the Travel Industry and its Potential in Improving Travel Accessibility
235	Yujie Wu		Big Data analytics in predict house price
236, 218	Weipeng Yang, Geng Niu		Big Data and Its Application in Education
237	Tousif Ahmed		Analyzing everyday challenges of people with visual impairments
301	Gagan Arora		Importance of Big data in predicting stock price
302	Sushant Athaley		Recipe Ingredients Analysis
304	Ricky Alan Carmickle		How Far have Space Walks Walked
hid305	error: yaml		How Far have Space Walks Walked
306, 338	Murali Cheruvu, Anand Sriramulu		Predicting Housing Prices - Kaggle Competition
308	Pravin Deshmukh	TBD	
hid309	error: yaml	TBD	
310	Kevin Duffy		Gerrymandering Detection Using Data Analysis
311	Matthew Durbin		Big Data in Genomics and Medicine
312	Neil Eliason		Big Data Mental Health Monitoring - A Private and Independent Approach
313	Tiffany Fabianac		The Impact of Clinical Trial Results on Pharmaceutical Stock Performance
314	Sarang Fadnavis	TBD	
315	Garner, Jeffry		TBI - A Data Driven Journey Beyond Contact Sports... Putting Data In The Drivers Seat
316	Robert Gasiewicz		Big Data and Edge Analytics in Weather Monitoring and Forecasting
318	Irey, Ryan		None
319	Mani Kumar Kagita		Face Detection and Recognition Using Raspberry Pi Robot Car
320	Elena Kirzhner		Big Data Applications in Real Estate Analysis
321	Knapp, William	TBD	
323	Uma M Kugan		None
324	Ashok Kuppuraj		Big Data Analytics in factors affecting Bitcoin
325	J. Robert Langlois		The importance of data sharing and replication, but what about data archiving?
326	Mohan Mahendrakar		None
327	Paul Marks		How Big Data Will Help Improve People's Health Worldwide
328	Dhanya Mathew		Predicting Profitable Customers in Banking Industry
329	Ashley Miller		Big Data and The Customer Experience Journey
330	Janaki Mudvari Khatiwada		Big Data Analytics in Monitoring Outdoor Air Quality
331	Tyler Peterson		Big Data Applications in Predicting Hospital Readmissions
332	Judy Phillips		Big Data Analytics to Reduce Health Care in the United States
334	Peter Russell		The Intersection of Big Data and IoT
335	Sean Shiverick		Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics
336	Jordan Simmons		None

337, 333	Ashok Reddy Singam, Anil Ravi	IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)
339	Hady Sylla	Diagnosis of Coronary Artery Disease Using Big Data Analysis
340	Timothy A. Thompson	New Approaches to Managing Metadata at Scale in Research Libraries
341	Tibenkana, Jacob	Not submitted
342	Nsikan Udoyen	TBD
343	Borga Edionse Usifo	Income Prediction Using Machine Learning Techniques
345	Ross Wood	Agricultural Data Science
346	Zachary Meier	Big Data Analysis for Wild File Prevention and Tracking
347	Jeramy Townsley	Killings by Police in the United States
348	Budhaditya Roy	Big Data Application in Precision Medicine and Pharmacogenomicsn

# Big Data Analytics in E-commerce

Himani Bhatt, Mrunal L Chaudhary

Indiana University

Bloomington, Indiana

himbhatt@iu.edu,mchaudh@iu.edu

## ABSTRACT

Humongous amounts of data gets generated every day in the domain of E-commerce industry. With the increasing competition and ever-changing market trends, it is a challenging task for the store owners to strategize business and marketing activities. If the companies are able to predict customer behavior, they can come up with business designs which can help them in making predictions about the customer purchasing patterns and thereby increase their revenue. In this project we have aimed to do analysis on the data of an E-commerce non-store online retail giant based in UK. The dataset, available in the UC Irvine repository by the name of 'Online Retail', consists of the goods purchased by different customers at a given time. Through this data available to us, we have done customer segmentation on the basis of the type and amount of goods purchased by a customer. We achieved this by doing a thorough exploration of the data, data pre-processing and then running different Machine Learning Classifiers to classify the customers in different categories.

## KEYWORDS

HID 202, HID 205, i523, Machine Learning, Analysis, E-commerce, retail, Customer Segmentation, Python, Regression, Boosting, KNN, Random Forest.

## 1 INTRODUCTION

The E-commerce industry is in constant shifts due to the ever-increasing changes in the technologies used to develop and maintain the E-commerce systems, the services that they are willing to offer, the market strategies which gain popularity at the time, and most importantly- the customer behavior [3]. The online store owners are the ones who are most affected by these changes. And since the competition in the field of E-commerce is fierce, the online store owners need to come up with business strategies and technologies which provide better customer services leading to their satisfaction and earning customer loyalty. To achieve this, they need to address these ever changing issues to survive and thrive in the E-commerce market and come up with better decisions faster. The key to achieving this lies in better understanding of the customer behavior and their purchasing patterns. That is where analytics comes into play. Analysis of customer behavior and purchasing patterns helps in devising better and accurate marketing strategies which can not only help in generating more profits but also in saving both time and efforts that goes into trying and testing different marketing activities [3]. This ability to capture and analyze user data, and then provide useful and in depth insights in it is what Machine Learning empowers us with. In this project, we aim to do analysis on a data set 'Online Retail' from the UC Irvine Machine Learning Repository to determine the customer purchasing pattern by using

different machine Learning algorithms like K-Means Clustering, Logistic Regression, Random Forest, Gradient Boosting, etc.

## 2 BACKGROUND

Before the advent of the World Wide Web, transactions that happened on a day to day basis meant physical presence of customers, the brick-and-mortar setting of a store which offered a limited variety of goods. With the evolution of internet and its application in retail, the field of E-commerce emerged and changed the entire facet of shopping. Since a proper set-up of a store is no longer needed, customers can buy goods at much lower prices, with a wider variety to choose from and that too without the need of physical presence. The online market is expected to grow by almost 56% from the year 2015 to 2020 [16]. In the United States alone, 56% of the population prefers to shop online. The E-commerce industry is growing at an average rate of 23% every year, with 90% of the Americans having done online shopping at some point in their lives [15]. With so many transactions happening over the internet, naturally the amount of data getting generated is humongous. Also, with the constantly changing market trends, strategies to overcome the competition and make profits need to be constantly improved. The key issues therefore are managing the data and drawing insights from them which will help in bettering the business decisions. To store and maintain the magnanimous amounts of data getting generated *everyday* is a huge hassle, because along with the volume, this data gets generated at a break neck speed and in different formats from traditional numeric databases to unstructured text documents [12]. The big data technologies like Hadoop and Spark can be used in addressing these hurdles, namely the volume, velocity and variety.

### 2.1 The Three V's of E-commerce Big Data

Like other technologies which deal with a humongous amount of data, E-commerce must also respond to the 3 Vs, namely Volume, Velocity and Variety:

**Volume** Thousand of online transactions happen every day making the data generation a real time process. The integration of Big Data involves collection of relevant data like customer behavior statistics on the basis of their searches, transactions, demography, etc. The challenge here is not only gathering the data but also in analyzing it.

**Variety** The data from online transactions comes in different varieties, right from structured databases to unstructured text documents, videos, feedback emails and comments, and others. The retailers need to understand this for making the right business decisions by keeping a leeway for possible data fluctuations such as seasonal ad peak loads like Black Friday sales.

**Velocity** Handling the huge amounts of data which is generated at unprecedented rates is another challenge that needs to be taken care of. It is therefore imperative to do rapid analysis so that timely actions can be taken to sustain in the competition and boost the profit margins

Storing and maintaining the big data is a hassle in itself, but it will provide little value if proper analysis is not done on it. That is where we will be focusing on in this project - making sense of the data. Hence for the scope of this project, we have performed analysis on a small data set of around 45 MB. Machine Learning Algorithms learn from the data. Since we will be using Machine Learning Algorithms, the accuracy of Analysis will only increase with increase in the size of the dataset. We will be discussing this in further detail in the coming sections.

We have now established the fact that the E-commerce companies have a lot of data at their fingertips. Making use of this data is where the challenge lies. Machine learning is an approach by which insights can be drawn from digital data at a rate much faster than any human is capable of doing [7]. Following are some of the biggest challenges that are faced in the field of E-commerce which Machine Learning addresses successfully:

(1) Optimization of the Prices:

Pricing, and in that, online pricing is critically important. Since prices of the competitors are only a few clicks away, it is far easier for the customers to compare prices. Setting up the optimum price, by considering many factors like the prices set by the competitors, the time of the day, the type of the customer and the product's demand therefore is a difficult task. Machine Learning technology can set the prices by considering all these factors at once.

(2) Fraud detection:

The E-commerce industry, like the other industries, is susceptible to fraudulent activities. The consequences of these activities can lead to tarnishing the name of the company forever. Machine Learning helps in detecting and preventing the frauds by processing the repetitive data at a high speed.

(3) Search Ranking:

Machine Learning is capable of pulling information from patterns of search and purchase by considering the factors like preferences, content and search items and come up with a powerful search engine that shows what the customer exactly wants.

(4) Product Recommendations:

Machine Learning is capable of effortlessly quantifying the buying patterns of the customers and developing a recommendation engine which makes relevant product suggestions to them.

(5) Customer Segmentation and Personalization:

In any business, Customer base is the most important factor and therefore providing a satisfactory customer experience is of utmost importance. The biggest challenge that E-commerce systems endeavor to overcome is the separation from their customers. In person, a salesperson can quickly take in what the customers are saying, their economic status, their body language, and behavior to help them find better or desired products. The salesperson thus is able to *segment* customers, and provide them with a *personalized* shopping experience. With online shopping, it is very difficult to make this happen since an in depth understanding is needed of the vast amount of the data to provide tailored choices to the customers,

which can result in sale loss.

Machine Learning makes the biggest impact by making it possible to give personalized customer experiences which can boost the sales and thereby increase the revenue.

The type of analysis and Machine Learning Algorithm to be chosen depends solely on the data at hand. The data set we aim to analyze is a transnational data set that has been archived in the UC Irvine Machine Learning Repository under the name 'Online Retail'.

A thorough Exploratory Data Analysis on this data lets us know what kind of Machine Learning Algorithm needs to be used. Also, several models can be applied and the one which gives the best accuracy and precision against the test data can be chosen. To add icing to the cake, we can even combine the results given by the different models to make an ensemble model which gives an accuracy that is better than that of the individual algorithms. Machine Learning Algorithms are mostly classified as supervised and unsupervised Learning algorithms.

In Supervised Learning, each example is a pair of an input object and the corresponding output value, also called the supervisory signal. A supervised learning algorithm analyzed the training data to produce an inferred function which is used to map new, unknown output-value examples [8]. Since there is the output value to *supervise* the learning algorithm, such approach is called 'Supervised Learning'. The most commonly used Supervised Learning Algorithms are Logistic and Linear Regression, Bagging and Boosting Algorithms, Decision Trees and Random Forest. The Logistic Regression algorithm determines the relationship between the input and the output variables and generates a classifier model to predict the category to which a new example belongs to. Thus Logistic Regression is a classification algorithm [2]. Decision Trees are non parametric Supervised Learning Algorithms which create a model by learning simple decision rules inferred from data attributes to predict the value of a target variable. Decision Trees can be used either for Classification or Regression [5]. Bagging is a technique used to reduce the variance in the predictions by combining the result of multiple classifiers modeled on different sub-samples of the same data set. One of the most commonly and widely used implementation of Bagging is Random Forests. In Random forest, there are multiple trees which classify a new sample based on the set of attributes and a new sample is classified to that class which received the maximum 'votes' from the individual trees. In case of doing Regression with the help of Random Forest, the average of the outputs given by different trees is taken [10].

In Unsupervised learning, only the input data is known with no knowledge of the corresponding output variable. The goal therefore of the Unsupervised Learning Algorithms is to model the underlying distribution or structure in the data to understand the data more. Since there is no output available to validate or 'supervise' the answers, such learning algorithms are called Unsupervised. The most common application of unsupervised learning is clustering. Clustering enables to differentiate the data by discovering the inherent groupings of the input. The most common implementation of Clustering is the K-means algorithm. This algorithm works iteratively to assign each data point to one of the k-groups based on the feature similarity.

### 3 EXPLORING THE DATASET

The dataset taken for the analysis is the ‘Online Retail’ data set available on the UCI Machine Learning Repository. This is a transnational dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data set consists of 5,41,909 transactions and 8 features which describe each of these transactions. There are missing values present in the dataset. All the attributes are integer and real numbers. The size of the dataset is 43.4 MB.

#### 3.1 Attribute Information

**InvoiceNo** : This refers to the Invoice number. It is a Nominal, 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.

**StockCode** : This refers to the Product (item) code. It is a Nominal, 5-digit integral number uniquely assigned to each distinct product.

**Description** : This refers to the Product (item) name. It is of Nominal data type.

**Quantity** : This refers to the The quantities of each product (item) per transaction. It is Numeric in type.

**InvoiceDate** : This refers to the Invoice Date and time. It is Numeric, and represents the day and time when each transaction was generated.

**UnitPrice** : This refers to the Unit price. It is of Numeric type, and represent the Product price per unit in sterling.

**customerID** : This refers to the Customer number. It i a Nominal, 5-digit integral number uniquely assigned to each customer.

**Country** : This refers to the Country name. It is of Nominal type, and represents the name of the country where each customer resides.

## 4 DATA PREPARATION

### 4.1 Installation Steps

The project has been implemented in Python 2.7 version and we have used the Jupyter Notebook App for the program execution. The Jupyter Notebook Application is an application having server-client architecture which allows editing and executing notebook documents through a web browser. A notebook document is a human readable and machine executable document which can be executed for implementation of data analysis. The Jupyter Notebook Application can be executed on the local host or can be installed on a remote machine accessed via the internet [6].

The Jupyter Notebook can be installed very easily on a machine which has either Python 2 or Python 3 version. Since we have implemented our project in Python 2.7, following commands are to be run in the terminal:

```
pip install --upgrade pip
```

The above command will upgrade the Python package manager (pip). 

```
pip install jupyter
```

The above command will install Jupyter in the local machine.

Once the Jupyter Notebook has been installed, it can be run using

the following command in the terminal:

```
jupyter notebook
```

This command will run the Jupyter notebook in the default browser of the machine on the default port 8888 of the localhost.

### 4.2 Packages Installation

Before running the code the following packages were imported/installed in the Python environment.

**Pandas** Pandas provide a very fast and flexible data structures to make working with relational data easy and fairly intuitive.

**Numpy** This is a fundamental package for scientific computation with Python and can be used as an efficient multi-dimensional container of generic data.

**Sklearn** Scikit-learn makes a wide range of supervised and unsupervised machine learning algorithms available in python. We have implemented all the Machine Learning Algorithms using this library.

### 4.3 Null Value Treatment

Before going ahead with Data Exploration, a quick look through the data showed many missing values. Hence before doing any analysis, it is imperative to treat the missing values. The dataset has almost 25% of the entries that are not assigned to any of the customer i.e. customerID attribute for those entries is null.

Missing value treatment can be done by deleting the columns and/or rows which have missing values beyond a decided threshold, or replacing them with the attribute mean, median or mode. Since the missing values in our case is the customerID, the replacement method cannot be applied. Also, these entries are useless for the analysis since we aim to do Customer Segmentation and without knowing the customerID, it cannot be achieved. Hence we have deleted the rows with missing customerID. After removing these entries , the dataset left is with 4,06,829 transactions.

The content of the dataset appears as shown in the Figure 1.

[Figure 1 about here.]

We also removed the duplicate values present in the dataset. There are 5225 such entries present in the data set that are deleted.

## 5 EXPLORING THE CONTENT OF VARIABLES

The dataframe has 8 variables and we can draw some inferences by analyzing these variables.

### 5.1 Countries

From the data we can see that there are 37 different countries from which orders were placed. We can determine the number of orders per country by a ‘Chloropeth’ map. A Chloropleth map shown in Figure 2 uses different colors and shades within predefined areas to indicate quantities in those areas.

[Figure 2 about here.]

The Figure 2 shows that maximum number of orders are placed from UK.

## 5.2 Customers and products

On observing the number of users, products purchased and number of transactions made; we can see that these are not proportional. This suggests that there were many transactions made for cancelling the orders shown in Figure 3

[Figure 3 about here.]

We can also determine the number of products purchased in each transaction. It shows that some customers purchased goods in bulk whereas some purchased a single product in a transaction. Also the orders with InvoiceNo starting with C are the cancelled orders. The details are shown in Figure 4.

[Figure 4 about here.]

## 5.3 Cancelled Orders

Almost 16% (3654) of the transactions are corresponding to the cancelled orders. In the dataset, corresponding to each cancelled transaction we should have an order placed with same quantity of products requested. While checking the same in the dataset, we found the details shown in Figure 5 for some of the orders.

[Figure 5 about here.]

This hypothesis should apply to the complete dataset, but on checking the whole dataset it is found out that there are some cancelled orders without the purchase order (the history of the order) made. This is done by locating the entries that indicate a negative quantity and then checking if there is an order indicating the same quantity (but positive) with the same description and the same customerID. We still get negative quantities. Going deeper in to this suggests that the entries with description 'Discount' have negative quantities associated with that transaction. And hence, to do the verification, we eliminated the 'Discount' entries. But again the initial hypothesis do not match; we still have negative numbers appearing in the quantity.

This can be because the buy orders were performed before December 2010 (the point of entry of the database). We can delete the records where a cancel order exists without the corresponding purchase order or where there is at least one counterpart with the exact quantity (since both records are logically cancelling each other). Total 8795 such records are found and deleted from the dataset.

## 5.4 StockCode

The StockCode variable should ideally contain letters. So we have filtered out the codes with only letters. We can observe from Figure 6, different type of transactions based on these (example D is for discounted transaction).

[Figure 6 about here.]

## 5.5 Basket Price

We have added a new variable to indicate total price of the purchase (by multiplying unit price of each product with quantity purchased). Each transaction corresponds to the prices for a single product. On

grouping the records based on a single order, we can see the complete price for that order as shown in Figure 7.

[Figure 7 about here.]

We can visualize the orders distinguished on the basis of total price of the basket. It can be shown as Figure 8 using a pie-chart.

[Figure 8 about here.]

It shows that majority of the orders are the bulk purchases since 60% of the orders have amounts greater than 200 Sterling.

## 6 EXPLORING PRODUCT CATEGORIES

The dataset contains two variables- Stockcode and Description defining products. We can categorize the products based on the content of the description variable. This can be done in the following way. Firstly, the proper or the common names appearing in the products' description are extracted. Then the root of the word and combining set of names associated with this particular root is extracted. Lastly, the frequency of the word is found in the description variable of the dataframe.

Upon checking, we found that there are 1483 keywords present in the description variable of the dataset. The most common keywords can be determined based on the occurrences. The Figure 9 shows the top word occurrences.

[Figure 9 about here.]

### 6.1 Categorizing Products

We have obtained around 1400 keywords from the above occurrence list , most of which do not make sense. After discarding the keywords that are appearing less than 13 times, we are left with 193 keywords that we will consider for our analysis.

These significant keywords are used for creating categories of the products. The data has been encoded using the principle of one-hot-encoding.

**One hot encoding** - One hot encoding is a process by which categorical variables are converted into a binary format of 0's and 1's that could be provided to ML algorithms to do a better job in prediction. The words present in the descriptions of the products are encoded. Also price range column is added as it will help in balanced grouping of the products.

### 6.2 Clustering of products

In the previous step we have created a matrix with encoded version of words present in the description variable. K-means clustering is used for the cluster assignment and since the data is in binary format because of encoding, the most appropriate distance method will be Hamming's metric (other distance functions are euclidean distance, Manhattan distance, binary distance, etc). It basically measures the minimum number of substitutions required to change one string into the other. But since the k-means package available in sklearn uses Euclidean distance by default, we have used it for our analysis.

Selection of optimum K-value:

The number of clusters can be selected using silhouette analysis on K-means clustering. It is used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to the points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

The Figure 10 shows silhouette score for different values of k. These scores do not have significant differences, but since for k value greater than 5, the resulting clusters have very few elements in them, we have taken k as 5.

[Figure 10 about here.]

### 6.3 Validating Quality of Classification

**6.3.1 Silhouette Score.** From the silhouette plot shown in Figure 11 we can see that cluster 1 has more number of elements than the other clusters. But overall distribution of elements in the clusters is comparative. Same can be seen from the Figure 12.

[Figure 11 about here.]

[Figure 12 about here.]

**6.3.2 Principal Component Analysis.** The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The initial matrix has large number of variables and hence, PCA is used for dimensionality reduction. From the Figure 13 we can say that we need more than 100 components to explain 90% of the variance in the data.

[Figure 13 about here.]

Another application of PCA is that it sets the indication of flicterfi membership. Biplot is the best example that can be provided here to support this idea. Using biplot, we get the indication of number of clusters in a dataset. Below Figure 14 shows these on limited number of components (since it is only for visualizing cluster distribution). We can observe the groupings of points or clusters as expected.

[Figure 14 about here.]

## 7 EXPLORING CUSTOMER CATEGORIES

In the previous section, we have divided products in 5 clusters. We have added a dummy variable categ\_product to indicate the cluster to which that customer belongs. Based on the clustering done on products we have created variables categ\_0..4 which stores amount spent on each of the product category. And the categ\_product variable which we have just created will have initial cluster assignment

based on these variables. These can be further grouped on the basis of InvoiceNo as shown in Figure 15.

[Figure 15 about here.]

### 7.1 Subsetting dataframe based on Time

We have taken 12 months data for the analysis. This can be done on the basis of variable InvoiceDate present in the dataset. Using this data we have developed a model to characterize and anticipate the habits of customers using the site and this, we are doing it from the first visit.

In the previous section we have seen the basket price of each invoices. For further analysis we will combine these on the basis of customerID to analyze the number of purchases made by each customer as shown in Figure 16. A customer category of particular interest is that of customers who make only one purchase. So one objective may be, for example, to target these customers in order to retain them. In the dataset we have almost one-third of the customer base similar to this.

[Figure 16 about here.]

### 7.2 Categorizing Customers

The information transactions per user is used for characterizing different types of customers. Because of different ranges of variations of different variables we have first scaled the data set. As done in the case of product categorization, we have again used K-means algorithm for cluster assignment.

Using the silhouette score, the optimum value of k comes out to be 11. The assignment of customers into different clusters is shown in Figure 17

[Figure 17 about here.]

Now we will check validity of the cluster assignment using PCA and Silhouette plot as done in the case of product categorization.

**7.2.1 PCA.** There is a certain disparity in the sizes of different groups that have been created. So we have validated it using PCA. From the representation shown in Figure 18, it can be seen, for example, that the first principal component allow to separate the tiniest clusters from the rest. More generally, we see that there is always a representation in which two clusters will appear to be distinct.

[Figure 18 about here.]

**7.2.2 Silhouette Plot.** As with product categories, another way to look at the quality of the separation is to look at silhouette scores shown in Figure 19 within different clusters:

[Figure 19 about here.]

We can see that the different clusters are indeed disjoint.

## 8 CLASSIFICATION OF CUSTOMERS USING CLASSIFICATION ALGORITHMS

In the previous section, we have made different client categories. In this part we will adjust a classifier so that the consumers can be classified in different client categories. The main aim of this is to enable the Classification on the first visit of the customer. To do this, we have defined a class that will allow interfacing the common functionalities to the different classifiers. Since we are going to classify the client on the basis of his/her first visit, the only parameters that we take into consideration are the contents of the basket and not the frequency of visits or the variation in the basket price over a period of time. Once this is done, we have split the dataset into train and test sets. The classification algorithms which we used to do this are mentioned below.

Before we delve deeper into the Classification Algorithm, some important concepts that need to be addressed are Cross Validation, Bias, Variance, underfitting and overfitting of the model.

**Variance** Variance essentially means how much the models estimated from the different training sets differ from each other. It measures how much the predictions made for a given point vary between the different realizations of the model [4]. When the training data tries to fit all the sample points to define the model, even the outlier data points, which are nothing but the noise, affect the model. Usually, the variance increases with increase in the complexity of the model.

**Bias** Bias essentially means how much the average model over the training sets differ from the true model. Bias usually occurs if the model is over simplified or if some inaccurate assumptions are made. Thus Bias increases with increase in the over simplification of the model [4].

**Underfitting** Scientific study of mental processes and behaviour.Underfitting occurs when the model is too simple to make relevant classification of the testing data [4].Thus when a model possesses high bias and low variance, we say there is underfitting of the model.

**Overfitting** Overfitting of a model occurs when the model is too complex and tries to fit in the irrelevant/outlier datapoints from the training set which is nothing but the noise [4]. Thus when a model possesses low bias and high variance, we say there is overfitting of the model.

**Cross Validation** Cross Validation is a technique for evaluating the predictive models by partitioning the original dataset into a training set to train the model, and a testing set to evaluate the model [11]. There are different ways to implement Cross Validation, the most effective of them all is the K-fold Cross Validation. In this method the dataset is divided into k subsets, out of which one is used as the test data and the remaining  $k - 1$  are combined together to form training data. This process is done k times, ensuring that every single sample in the dataset gets to be tested exactly one time and gets trained upon exactly  $k - 1$  times. The variance therefore gets decreased as the k increases [11].

### 8.1 Logistic Regression

Logistic Regression as mentioned before is a Supervised Learning method which does analysis on a dataset containing two or more independent variables for determining the outcome. This outcome, i.e the dependent variable, is binary in nature, meaning it can have only two possible outcomes [9]. Multinomial Logistic Regression as the name suggests, generalizes the Logistic Regression to multiple classes, meaning the model can be used to predict the probabilities of the different outcomes of a categorically distributed dependent variable [17]. The goal of a Logistic Regression model is to determine a fitting model which best describes the relationship between the dependent variables (output variable) and a set of the input independent variables. Logistic Regression generates the coefficients along with the standard errors and significance levels of the below equation for predicting the logit transformation of the probability of presence of the characteristics of interest in a given sample example [9].

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + \beta_k X_k$$

where p is the probability of the presence of a characteristic of interest. and  $\text{logit}(p) = \log(p/1-p)$  In logistic Regression, the goal is to choose the parameters  $\beta$  in such a way that the likelihood of observing the new sample values is maximized [9].

In the Python code, we have imported the module ‘linear\_model’ from the ‘sklearn’ package to perform Logistic Regression by using the function ‘logistic\_regression’. And we have taken the  $k = 5$  for k-fold cross validation. While performing Logistic Regression, we created an instance of the Class\_Fit class and then ran the model on training data and see how the predictions are made as compared to the real values. The learning curve graph is as shown in Figure 20.

[Figure 20 about here.]

As we can see from the Figure 20, when the number of training examples increases, the cross-validation and train curves almost converge towards the same limit suggesting that the model has low variance. Thus we can say that model is not suffering from overfitting. Also one point to note is that the accuracy is high, which means that the model has low bias, thus suggesting that it does not under-fit the data. The precision which we got from running the Logistic Regression model on the training data is 88.78%.

### 8.2 K Nearest Neighbours

KNN is a non parametric algorithm which means that there are no underlying assumptions that are made on the data. Also it is a lazy learning algorithm meaning that it does not do any generalization by using the training data. All the training data is needed during the testing phase [13].

KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this is be the model (or most common) class value. To elaborate on this, KNN makes predictions using the training dataset directly. These predictions are made for a new sample by going through the entire training set to find k such samples which are most similar or which are the ‘neighbors’ of the the new instance. Once these k

instances are found out, the output variable corresponding to these is summarized and in case of Classification, it gives a class value to which the new instance belongs. The k ‘neighbors’, i.e., the most similar instances from the data set are found by using the distance measure- k such instances whose distance from the new instance is the least [13]. There are many distance functions which can be used, the most popular being the Euclidian distance function, the formula for which is given by:

$$\text{EuclideanDistance}(x, x_i) = \sqrt{\sum((x_j - x_{ji})^2)}$$

where  $x$  is a new data point and  $x_i$  is an already existing point [13].

The optimum value of K can be found by algorithm tuning, i.e. running the algorithm over several values of k and finding out and then figuring out for which k the algorithm gives the best results [14].

The output, i.e the class of the new sample can be calculated as the class which has the highest frequency from the k neighbours. Thus, each of the instances votes for their own class and the class which gets the maximum votes is taken as the prediction value [14].

In Python, the ‘neighbors’ library is imported from the sklearn package which performs the KNN classification through the KNeighborsClassifier function. The parameters that are used are ‘n\_neighbors’ which represents the number of neighbors to use, in our case we have used the np.arange method to give sequence from 1 to 49. Also, we run the model using the K-fold Cross Validation with the value of  $k = 5$ . Once the model is run, we have drawn the learning curve graph which is as represented in the Figure 21.

[Figure 21 about here.]

The precision which we got from running the KNN model on the training data is 80.33%.

### 8.3 Random Forest

As the name suggests, Random Forest is an ensemble classifier which consists of many classification trees. An ensemble classifier is a multiple classifier algorithms, decision trees in the case of Random Forests, and the final output is the combined output of the all the classifier algorithms. In our case we will be using Random Forest Algorithm for classification of the clients into different categories. A Random Forest grows many trees. For classifying a new object from an input vector, each tree in the forest gives a classification and vote for a particular class. And the forest then chooses the class having maximum number of votes over the other classes [1].

The question here that needs to be addressed is, how does the growth of a tree happen?

Each tree is grown as follows:

If the training set consists of N cases, then N cases are sampled with replacement from the original data. This is the training set for growing a tree. Thereafter, a number  $m \leq M$  which is the number of input variables is taken such that the best split obtained on these m is used to split the node. The value of m is constant throughout the forest-growing. Each tree is allowed to grow to the fullest possible extension [1]. In Python, the ‘ensemble’ library is imported from the sklearn package which performs the Random Forest classification through the RandomForestClassifier function. The parameters given to this function are criterion, n\_estimators and max\_features. The criterion is used to measure the quality of the split. The Gini is

for measuring the Gini impurity and Entropy is for information gain. The max\_features are the number of the features that can be chosen when looking for the best split. For ‘sqrt’, the number of maximum features chosen are square root of the number of the features and for ‘log’, it is log of the number of the features. And the n\_estimators is the number of trees in the forest. Once the model is run, we have drawn the learning curve graph which is as represented in the Figure 22 .

[Figure 22 about here.]

The precision which we got from running the Random Forest model on the training data is 90.17%.

### 8.4 Gradient Boosting Classifier

AdaBoost Classifier, short for Adaptive Classifier is another example of ensemble classifier. It is a general ensemble method which creates a strong classifier by combining the outputs of the weaker learning algorithms into a weighted sum to finally provide the output of the *boosted* classifier. This is done by building one model from the training set and then building a second one which attempts to rectify the errors from the first model and so on until either the limit of maximum models that can be added is reached or the training set is predicted accurately. AdaBoost is an adaptive algorithm, meaning that the weak learning algorithm can be tweaked to create a stronger classifier [6].

The Adaptive Boosting algorithm was recast into a statistical framework. “Arcing is an acronym for Adaptive Re-weighting and Combining. Each step in an arcing algorithm consists of a weighted minimization followed by a re-computation of [the classifiers] and [weighted input] [6]” This framework is called as Gradient Boosting.

Gradient Boosting involves three elements namely:

**A loss function** The selection of the loss function depends on the problem at hand. For example if it is a regression algorithm, then squared loss functions are used and if it is a classification algorithm then logarithmic functions are used. The main aim of the algorithm is to optimize the loss function.

**A weak learner** Regression trees are used as the weak learners in the Gradient Boosting Algorithm since they can output real values for splits which can be added together and the residuals in the predictions can be corrected. The weak learners are used for making predictions. These trees are constructed in a greedy manner usually up to 4-8 levels.

**An additive model** The additive model is made so as to add the weak learners to minimize the loss function. The trees are added one at a time with no changes to the existing trees in the model. A gradient descent model is used to reduce the loss when adding trees first by parameterizing the tree and then by modifying the parameters of the tree and moving in the right direction by reducing the loss in residuals. This approach is called Functional Gradient descent [6].

This framework was further developed by Friedman and called Gradient Boosting Machines. Later called just gradient boosting or gradient tree boosting [6].

In Python, the ‘ensemble’ library is imported from the sklearn package which performs the Gradient Boosting classification through the GradientBoostingClassifier function. The parameter given to this function is n\_estimators which is the number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number results in better performance. Learning curve is shown in the Figure 23.

[Figure 23 about here.]

The precision which we got from running the Gradient Boosting model on the training data is 90.86%.

Now that we have the results of all the models, we can combine them using VotingClassifier method that we imported from the sklearn package to improve the classification model. Since we have already found the best parameters for each of the classifiers, we have adjusted the parameters of the classifiers accordingly. So now the best parameters are taken and merged to define a classifier which we then trained on the data. When we created a prediction on this, we got the precision value as 90.44%.

## 9 TESTING THE PREDICTIONS

Until now we have done all the analysis on the data from the first 10 months. After this we test the model on the set\_test dataframe which contains the data of the last two months. The regrouping of the data is done according to the same procedure that we followed while regrouping the training data. But now we have to take into consideration the time difference in between the two datasets and the count(the total number of visits which the client made to the website) and sum(total amount that he/she spent) variables so that we have an equivalence in between the training set and testing set. The dataframe so obtained is now converted to a matrix and we retained only those variables that define the category to which the clients belonged. And just like on the training dataset the method of normalization was called, to maintain consistency, the same method is called on the test set as well.

Each row of the matrix obtained now represents the buying habits of the customers. Now all we have to do is to define the category to which the customer belongs by using these habits. The important point to note here is that this is just the test data preparation step by defining the category to which the consumer belongs for a period of two months through the variables count, min, max and sum. Thus this step *does not* correspond to the classification step itself. The classifier that we defined in the step 5 uses variables that were defined from the client’s first purchase.

So now, we have the data available for two months, and through that we can define the category to which the consumer belongs. The predictions now obtained by running the classifiers on test data can be tested against these categories. The instance of the k-means clustering method that we used in the Customer Categories section is used to define the category to which a client belongs. This contains the predict method which will calculate the distance of the consumers from the centroids of the 11 categories that we deduced, and the category which is closest to the clients’ buying habits will define his/her category. Thus all we need after this for the execution of the classifier is to select the variables on which it acts, i.e. on mean, cat\_0, cat\_1, cat\_2, cat\_3 and cat\_4. After examining the predictions of the different classifiers, we get precision scores as

shown in Table 1.

[Table 1 about here.]

And now, like we did in the Section 5, we will use the voting classifier method to merge the results obtained by these individual classifiers and see whether they combined result is better than the individual. It turns out that it is. We get the precision rate for the combined classifier to be 76.48% for the test data set. This concludes the analysis phase.

## 10 CONCLUSION

E-commerce is one of the emerging fields for Data Analysis since a lot of data gets generated every day at a break-neck speed in many different formats. To sustain in such a business, a very robust and extensive data analysis is needed to keep up with the ever changing markets by implementing different marketing strategies. And since the whole business revolves around the customers, they form the most important aspect of the analysis. We have tried to achieve Customer Segmentation on the basis of the purchasing patterns and frequency of client visits to their online portal. The dataset on which we performed analysis provided details on the purchases made by the consumers over a period of more than a year. Every entry in the dataset contained the purchase of a particular product on a given date by a particular customer. Out of the 591909 entries made in the dataset, approximately 4000 different consumers are present. From the information available for each consumer, we decided to go ahead with Customer Segmentation analysis by developing a classifier that predicts the type of purchase a consumer would make and his/her frequency of visits to the E-commerce website.

In the first step of this classification, we found out the different products sold by the company, and then classified the products into 5 categories of goods by using K-means clustering. In the second step we performed the classification of the customers on the basis of purchasing habits in the first 10 months. The customers were classified into 11 categories on the basis of the types of products they usually bought, the number of visits they made to the website and the amount for which they shopped over a period of 10 months. Once we had the categories of the consumers, we performed training of the data of the first 10 months using different classifiers namely Logistic Regression, Random forests, KNN and Gradient Boosting algorithms to classify the consumers in these 11 categories, on the basis of their first purchase. The classifiers were based on these variables: the total price of the current purchase and the percentage of the amount spent in each of the 5 product categories. Once the customers were classified in the 11 categories, the quality of the data set was tested on the remaining two months of the dataset. This was achieved in two steps. In the first step, we assigned the category to which each customer belonged to, and then the classifier predictions were compared against these categories. And then we combined the results of the various classifiers by using the Voting Classifier method. The model performed with a 76.48% of precision, that is 76.48% of the times the clients were awarded the right classes.

One bias which we did not consider while doing the analysis is the

seasonal fluctuations, like festive and seasonal sales. Since at these times the sales of products may rise and just before and after the sale duration, the sales may drop. Thus the purchasing habits of customers are dependent on the time of the year as well. Hence the seasonal effects may cause the actual sales in the last two months to be quite different from the ones which we extrapolated from the first ten months to the last two months. For overcoming such biases, it would be beneficial if the data were of a larger size and covered a larger period of time.

Knowing the type of a customer is critically important for an E-commerce business. By doing so, the store owners can provide personalized services to the customers, which will yield higher customer satisfaction. Customer satisfaction is directly proportional to the loyalty of the customers, thus Customer Segmentation and Personalization can help the company in increasing their brand name. Knowing the preferences and choices of customers also helps in catering to those needs of the customers which they may not be aware of in the first place. Thus by knowing the purchasing patterns of the customers, we can provide them with tailored suggestions, which can even increase the revenues of the company. Thus through proper implementation of the business strategies and marketing activities, which are motivated by a thorough Analysis of the data available can help the company in attracting loyal customers, increasing the revenue and establishing a better brand value.

## ACKNOWLEDGMENTS

The authors would like to thank Prof. Dr. Gregor von Laszewski for giving the opportunity to work on this project. The author would also like to thank the Associate Instructors of the class for their help and for answering questions on Piazza which helped everyone.

## REFERENCES

- [1] Jason Brownlee. 2016. <https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/>. (2016). <https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/>
- [2] Jason Brownlee. 2016. Supervised and Unsupervised Machine Learning Algorithms. (2016). <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>
- [3] Justin Butlion. 2015. An Introduction to Analytics for Ecommerce Websites. (2015). <https://blog.kissmetrics.com/intro-to-ecommerce-analytics>
- [4] Scott Fortmann-Roe. 2012. Understanding the Bias-Variance Tradeoff. (2012). <http://scott.fortmann-roe.com/docs/BiasVariance.html>
- [5] Prashant Gupta. 2017. Decision Trees in Machine Learning. (2017). <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
- [6] Antonino Ingargiola. 2015. What is the Jupyter Notebook? (2015). [http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what\\_is\\_jupyter.html](http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html)
- [7] Lacie Larschann. 2017. 7 Powerful Applications of Machine Learning in E-Commerce. (2017). <https://www.granify.com/blog/powerful-applications-of-machine-learning-in-e-commerce>
- [8] EILEEN MCNULTY. 2015. WHAT'S THE DIFFERENCE BETWEEN SUPERVISED AND UNSUPERVISED LEARNING? (2015). <http://dataconomy.com/2015/01/whats-the-difference-between-supervised-and-unsupervised-learning/>
- [9] Medcalc. 2017. Logistic Regression. (2017). [https://www.medcalc.org/manual/logistic\\_regression.php](https://www.medcalc.org/manual/logistic_regression.php)
- [10] Sunil Ray. 2017. Understanding Support Vector Machine algorithm from examples (along with code). (2017). <https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/>
- [11] Jeff Schneider. 1997. Cross Validation. (1997). <https://www.cs.cmu.edu/~schneide/tut5/node42.html>
- [12] Granner Smith. 2017. Big Data: Making It Big For E-Commerce Retailers. (2017). <http://www.digitalistmag.com/customer-experience/2017/04/28/big-data-making-it-big-for-e-commerce-retailers-05049637>
- [13] Saravanan Thirumuruganathan. 2010. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. (2010). <https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/>
- [14] Saravanan Thirumuruganathan. 2010. A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. (2010). <https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/>
- [15] Tracey Wallace. 2017. Ecommerce Trends: 147 Stats Revealing How Modern Customers Shop in 2017. (2017). <https://www.bigcommerce.com/blog/ecommerce-trends/>
- [16] Wikipedia. 2017. E-commerce. (2017). <https://en.wikipedia.org/wiki/E-commerce>
- [17] Wikipedia. 2017. Multinomial logistic regression. (2017). [https://en.wikipedia.org/wiki/Multinomial\\_logistic\\_regression](https://en.wikipedia.org/wiki/Multinomial_logistic_regression)

## LIST OF FIGURES

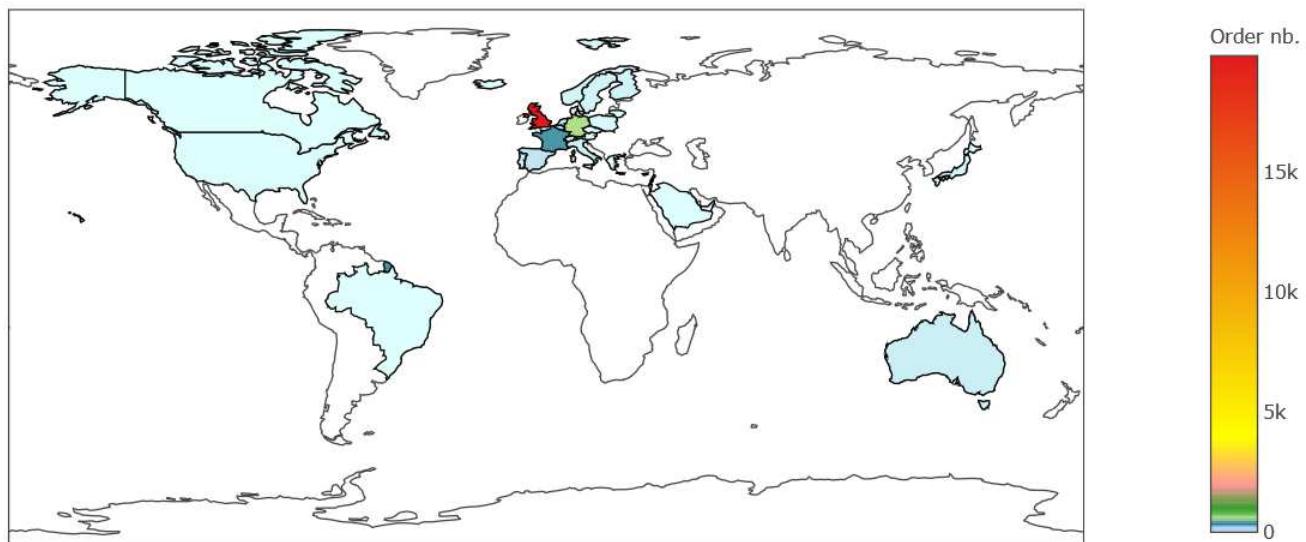
1	Data set Contents	11
2	Distribution of Orders based on Countries	11
3	Customer Products Transactions	11
4	Number of products per Customer	12
5	Transactions for Cancellation	12
6	Stock Codes	13
7	Basket Price	13
10	Silhouette Scores	13
8	Pie-Chart	14
9	Word Occurrences	15
11	Silhouette plot	16
12	Cluster Composition	17
13	PCA	18
14	Biplot	18
15	Table	18
16	Number of Purchases	19
17	Number of Purchase	19
18	PCA	19
19	Silhouette Plot	20
20	Logistic Regression Learning Curve	21
21	KNN Learning Curve	22
22	Random Forest Learning Curve	23
23	Gradient Boosting Learning Curve	24

**Figure 1: Data set Contents**

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

**Figure 2: Distribution of Orders based on Countries**

Number of orders per country



**Figure 3: Customer Products Transactions**

products	transactions	customers
quantity	3684	22190

Figure 4: Number of products per Customer

<b>CustomerID</b>	<b>InvoiceNo</b>	<b>Number of products</b>
0	12346 541431	1
1	12346 C541433	1
2	12347 537626	31
3	12347 542237	29
4	12347 549222	24

Figure 5: Transactions for Cancellation

	<b>InvoiceNo</b>	<b>StockCode</b>	<b>Description</b>	<b>Quantity</b>	<b>InvoiceDate</b>	<b>UnitPrice</b>	<b>CustomerID</b>	<b>Country</b>
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom
61624	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346	United Kingdom
286623	562032	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-08-02 08:48:00	4.25	12347	Iceland
72260	542237	84991	60 TEATIME FAIRY CAKE CASES	24	2011-01-26 14:30:00	0.55	12347	Iceland
14943	537626	22772	PINK DRAWER KNOB ACRYLIC EDWARDIAN	12	2010-12-07 14:57:00	1.25	12347	Iceland

Figure 6: Stock Codes

POST	-> POSTAGE
D	-> Discount
C2	-> CARRIAGE
M	-> Manual
BANK CHARGES	-> Bank Charges
PADS	-> PADS TO MATCH ALL CUSHIONS
DOT	-> DOTCOM POSTAGE

Figure 7: Basket Price

CustomerID	InvoiceNo	Basket Price	InvoiceDate
1	12347	537626	711.79 2010-12-07 14:57:00.000001024
2	12347	542237	475.39 2011-01-26 14:29:59.999999744
3	12347	549222	636.25 2011-04-07 10:42:59.999999232
4	12347	556201	382.52 2011-06-09 13:01:00.000000256
5	12347	562032	584.91 2011-08-02 08:48:00.000000000
6	12347	573511	1294.32 2011-10-31 12:25:00.000001280

Figure 10: Silhouette Scores

```
('For n_clusters =', 3, 'The average silhouette_score is :', 0.10158702596012364)
('For n_clusters =', 4, 'The average silhouette_score is :', 0.12680045883937879)
('For n_clusters =', 5, 'The average silhouette_score is :', 0.14553871352885445)
('For n_clusters =', 6, 'The average silhouette_score is :', 0.15122077520906058)
('For n_clusters =', 7, 'The average silhouette_score is :', 0.1463684372259042)
('For n_clusters =', 8, 'The average silhouette_score is :', 0.14764212603720744)
('For n_clusters =', 9, 'The average silhouette_score is :', 0.13974230402472737)
```

Figure 8: Pie-Chart

## Distribution of the amounts of orders

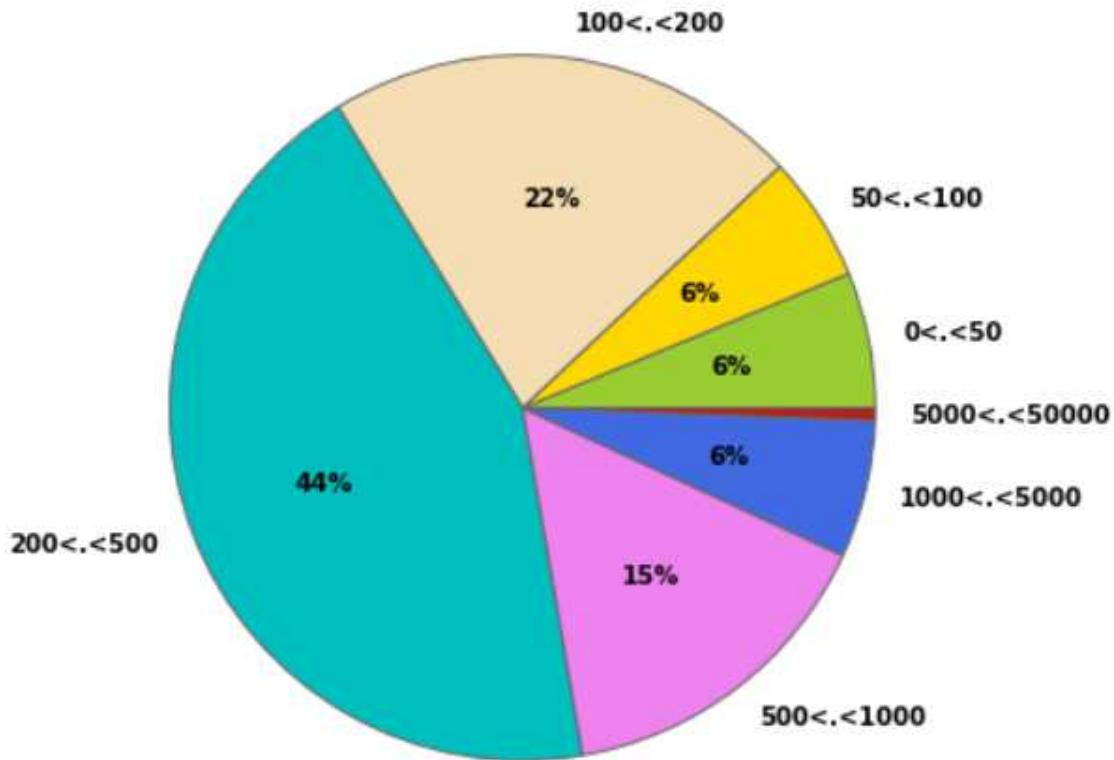


Figure 9: Word Occurrences

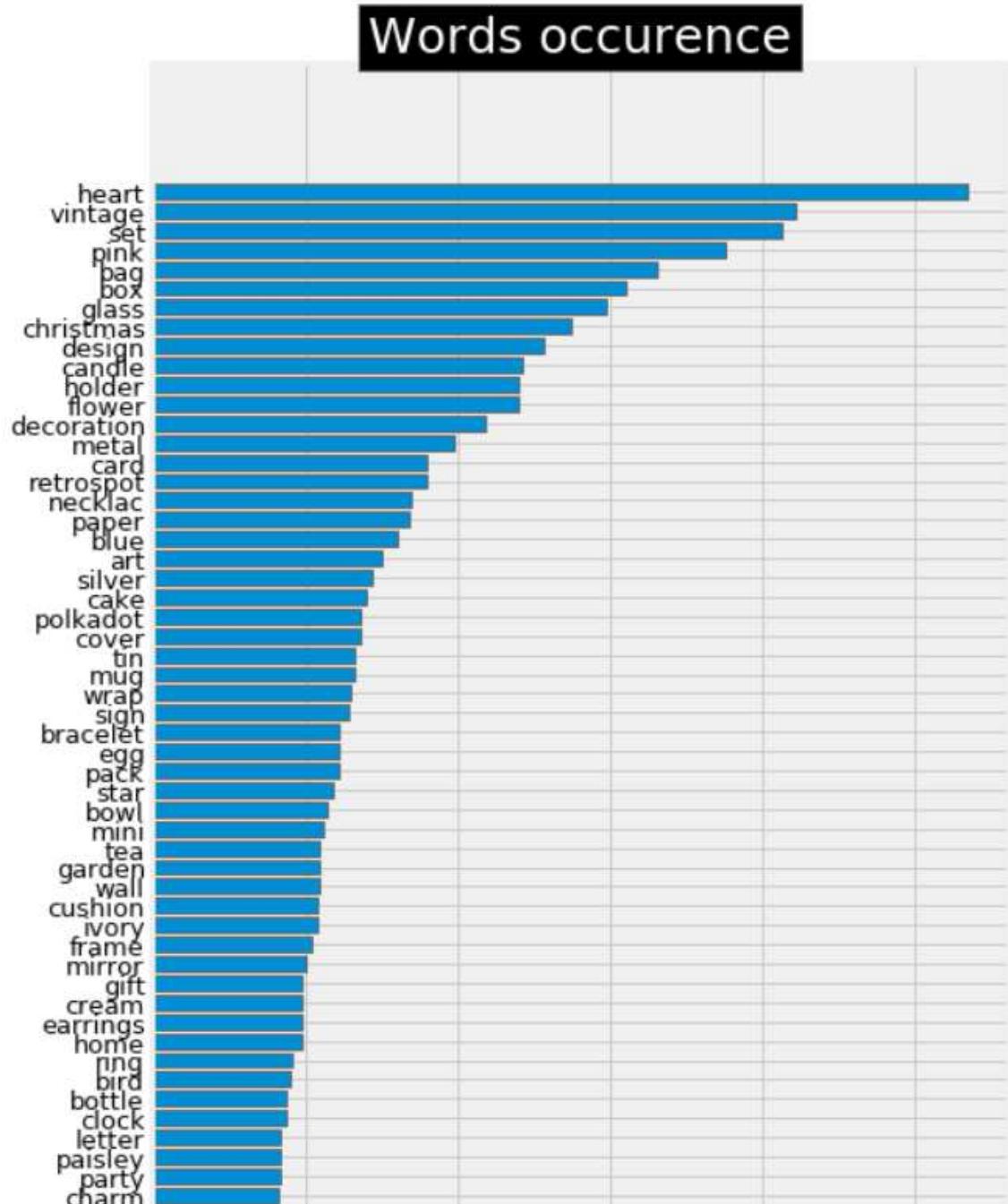


Figure 11: Silhouette plot

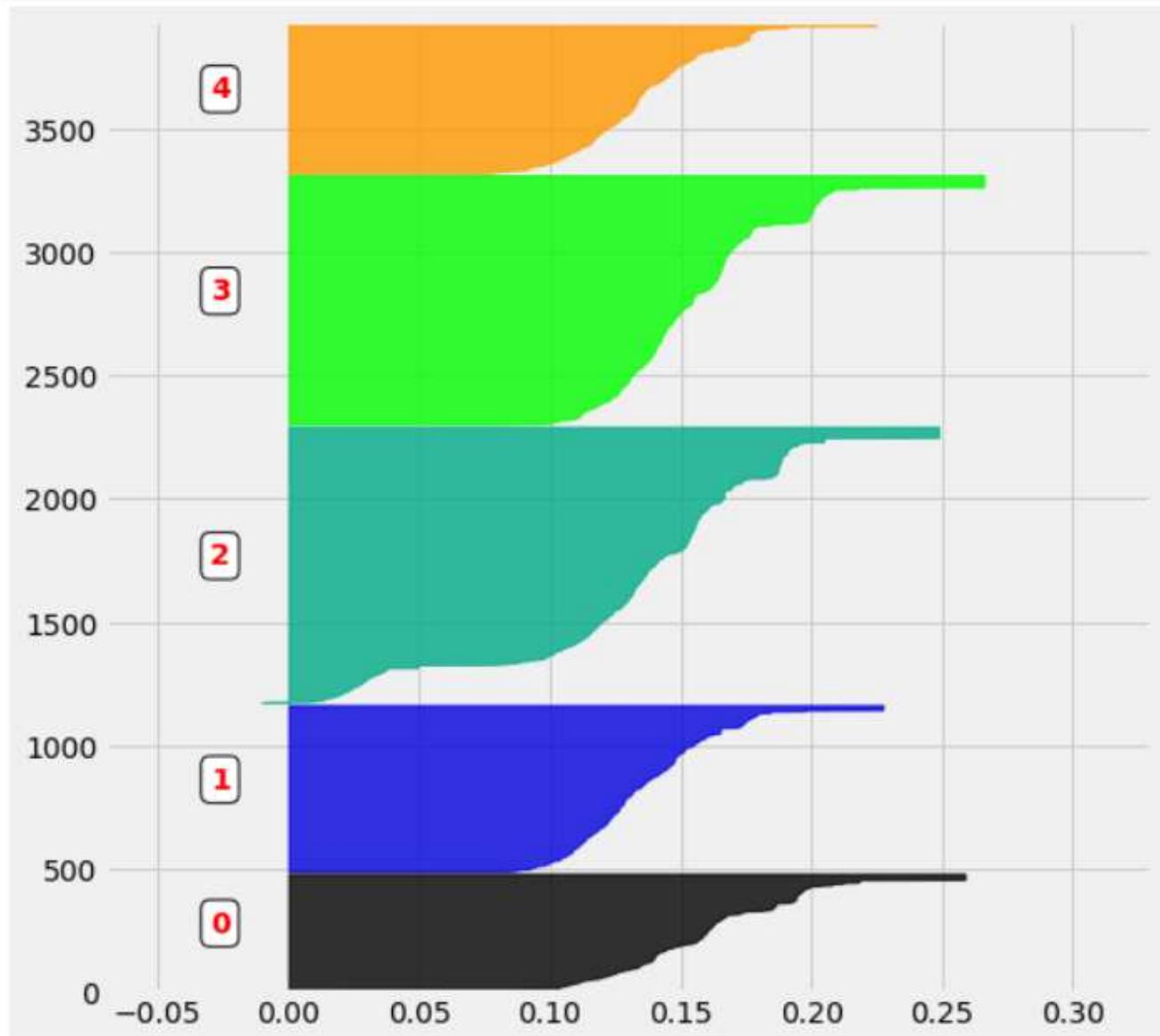


Figure 12: Cluster Composition

2	1118
3	1009
1	673
4	606
0	472

Figure 13: PCA

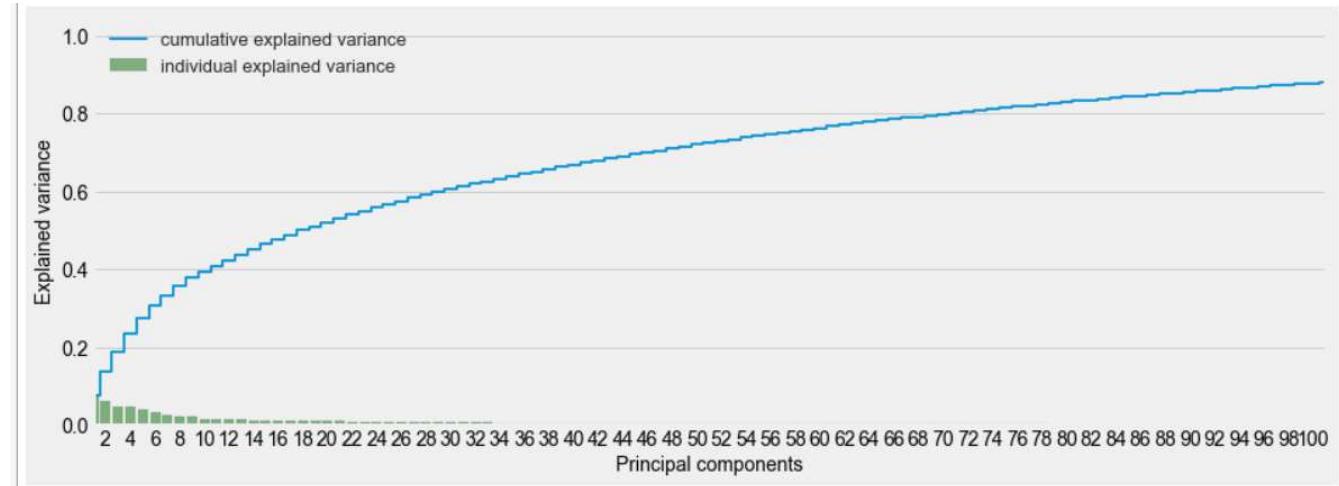


Figure 14: Biplot

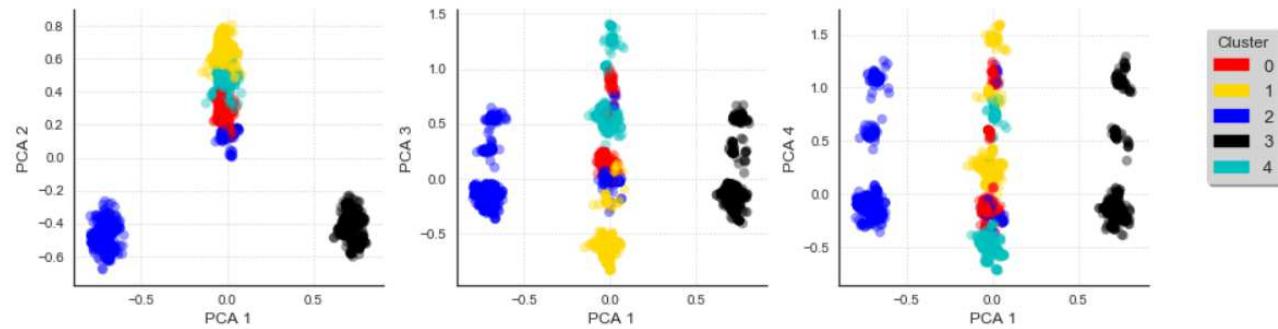


Figure 15: Table

	CustomerID	InvoiceNo	Basket Price	categ_0	categ_1	categ_2	categ_3	categ_4	InvoiceDate
1	12347	537626	711.79	124.44	83.40	23.40	187.2	293.35	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	0.00	53.10	122.59	130.5	169.20	2011-01-26 14:29:59.99999744
3	12347	549222	636.25	0.00	71.10	119.25	330.9	115.00	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	19.90	78.06	41.40	74.4	168.76	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	97.80	119.70	99.55	109.7	158.16	2011-08-02 08:48:00.000000000

Figure 16: Number of Purchases

	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_4
0	12347	5	382.52	711.79	558.172000	2790.86	8.676179	14.524555	14.554295	29.836681	32.408290
1	12348	4	227.44	892.80	449.310000	1797.24	0.000000	0.000000	58.046783	41.953217	0.000000
2	12350	1	334.40	334.40	334.400000	334.40	0.000000	27.900718	23.654306	48.444976	0.000000
3	12352	6	144.35	840.30	345.663333	2073.98	14.301006	3.370331	53.725205	12.892120	15.711338
4	12353	1	89.00	89.00	89.000000	89.00	22.359551	19.887640	44.719101	13.033708	0.000000

Figure 17: Number of Purchase

	1	0	2	6	3	9	8	7	5	4	10
<b>nb. de clients</b>	1484	451	371	344	296	284	191	161	9	9	8

Figure 18: PCA

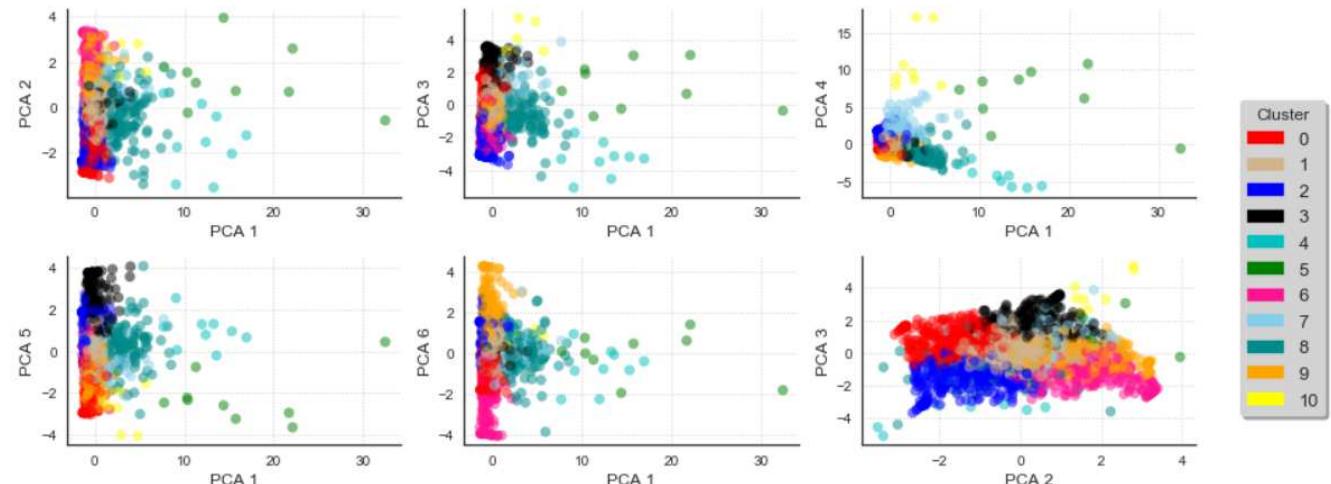


Figure 19: Silhouette Plot

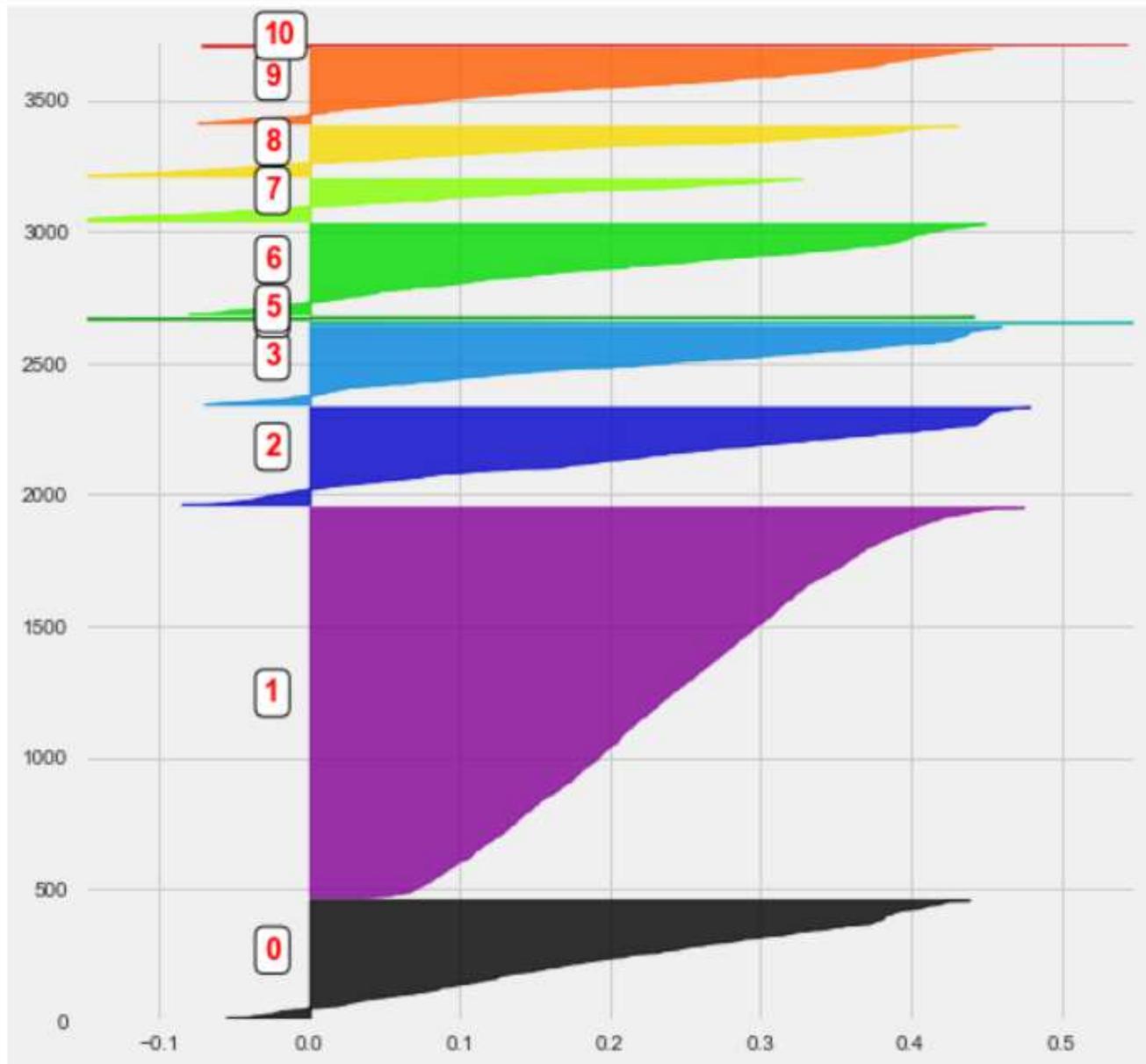


Figure 20: Logistic Regression Learning Curve

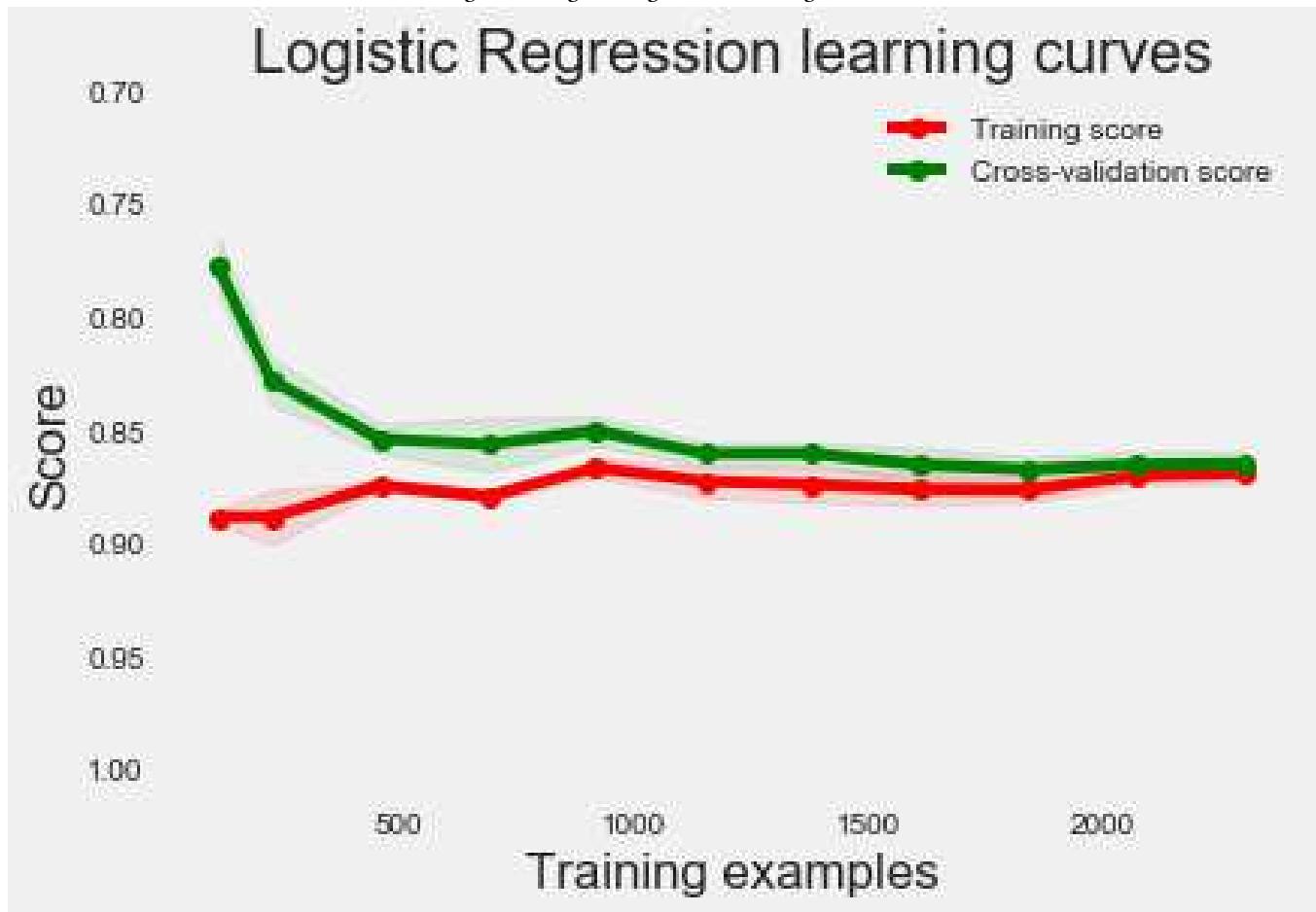


Figure 21: KNN Learning Curve

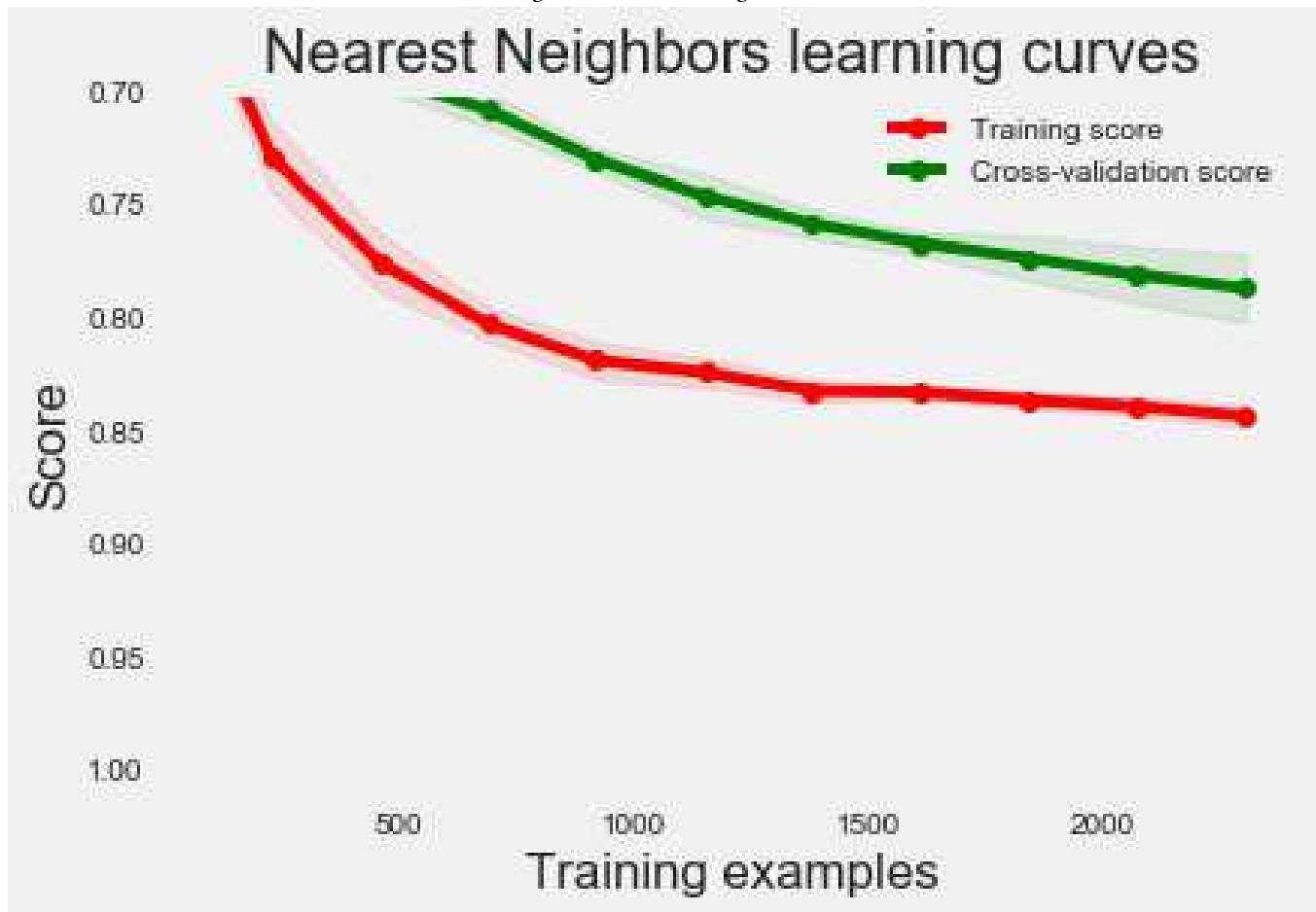


Figure 22: Random Forest Learning Curve

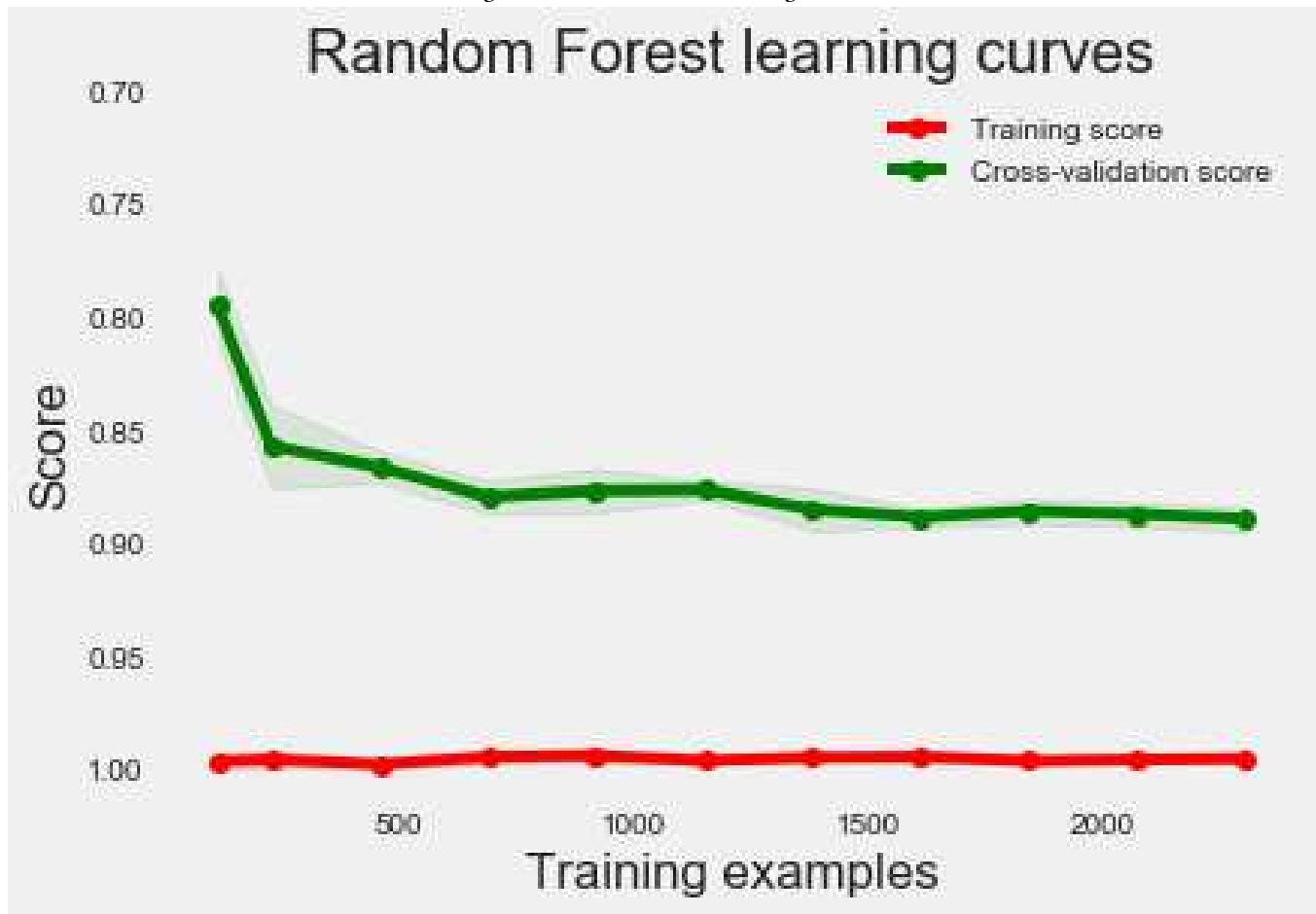
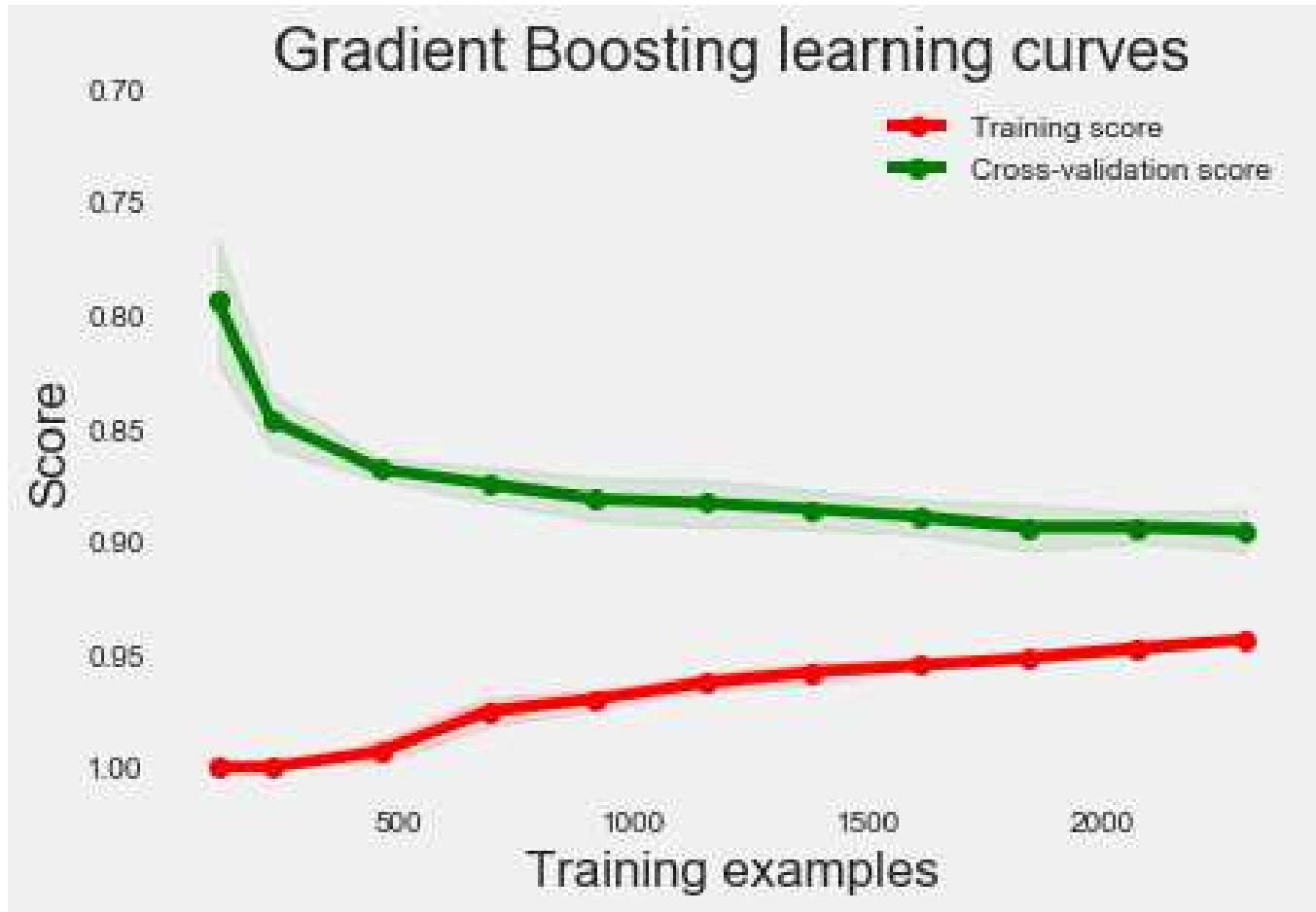


Figure 23: Gradient Boosting Learning Curve



LIST OF TABLES

1 Algorithms with their Precision Scores

26

**Table 1: Algorithms with their Precision Scores**

Algorithm	Precision(%)
Logistic Regression	72.99
KNN	68.44
Random Forest	75.93
Gradient boosting	75.74

## 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

```
[2017-12-10 13.47.18] 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.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.

Typesetting of "report.tex" completed in 1.2s.

./README.yml
30:14    warning  truthy value is not quoted  (truthy)
34:81    error    line too long (81 > 80 characters)  (line-length)
35:81    error    line too long (85 > 80 characters)  (line-length)
36:81    error    line too long (82 > 80 characters)  (line-length)
36:82    error    trailing spaces  (trailing-spaces)
37:81    error    line too long (82 > 80 characters)  (line-length)
38:79    error    trailing spaces  (trailing-spaces)
39:81    error    line too long (83 > 80 characters)  (line-length)
39:83    error    trailing spaces  (trailing-spaces)
```

```
40:81    error    line too long (82 > 80 characters)  (line-length)
40:82    error    trailing spaces  (trailing-spaces)
53:81    error    line too long (103 > 80 characters)  (line-length)
53:103   error    trailing spaces  (trailing-spaces)
54:81    error    line too long (104 > 80 characters)  (line-length)
54:104   error    trailing spaces  (trailing-spaces)
55:81    error    line too long (105 > 80 characters)  (line-length)
55:105   error    trailing spaces  (trailing-spaces)
56:81    error    line too long (107 > 80 characters)  (line-length)
57:81    error    line too long (110 > 80 characters)  (line-length)
58:81    error    line too long (111 > 80 characters)  (line-length)
59:81    error    line too long (116 > 80 characters)  (line-length)
59:116   error    trailing spaces  (trailing-spaces)
60:81    error    line too long (107 > 80 characters)  (line-length)
61:81    error    line too long (110 > 80 characters)  (line-length)
62:81    error    line too long (106 > 80 characters)  (line-length)
62:106   error    trailing spaces  (trailing-spaces)
63:81    error    line too long (110 > 80 characters)  (line-length)
68:27    error    trailing spaces  (trailing-spaces)
70:14    error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Himani Bhatt
hid: 202
paper1: Nov 04 17 100%
paper2: Nov 27 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
26
wc 202 project 26 8175 content.tex
wc 202 project 26 8196 report.pdf
wc 202 project 26 380 report.bib
```

```
find "
```

```
-----  
passed: True  
  
find footnote  
-----  
  
passed: True  
  
find input{format/i523}  
-----  
  
10: \input{format/i523}  
  
passed: True  
  
find input{format/final}  
-----  
  
passed: False  
  
floats  
-----  
  
112: The content of the dataset appears as shown in the Figure  
      \ref{data}.  
114: \begin{figure}  
116: \label{data}  
118: \includegraphics[width=\columnwidth]{images/DatasetContent.PNG}  
129: From the data we can see that there are 37 different countries  
      from which orders were placed. We can determine the number of  
      orders per country by a ‘Chloropeth’ map. A Chloropleth map  
      shown in Figure \ref{country} uses different colors and shades  
      within predefined areas to indicate quantities in those areas. \\  
131: \begin{figure}  
133: \label{country}  
135: \includegraphics[width=\columnwidth]{images/chloropleth.PNG}  
138: The Figure \ref{country} shows that maximum number of orders are  
      placed from UK.  
142: On observing the number of users, products purchased and number  
      of transactions made; we can see that these are not proportional.  
      This suggests that there were many transactions made for  
      cancelling the orders shown in Figure \ref{2.1}\\  
144: \begin{figure}  
146: \label{2.1}  
148: \includegraphics[width=\columnwidth]{images/2_1.PNG}
```

```

152: Also the orders with InvoiceNo starting with C are the cancelled
orders. The details are shown in Figure \ref{2.2}.
154: \begin{figure}
156: \label{2.2}
158: \includegraphics[width=\columnwidth]{images/2_2.PNG}
163: Almost 16\% (3654) of the transactions are corresponding to the
cancelled orders. In the dataset, corresponding to each cancelled
transaction we should have an order placed with same quantity of
products requested. While checking the same in the dataset, we
found the details shown in Figure \ref{2.3} for some of the
orders.\\
165: \begin{figure}
167: \label{2.3}
169: \includegraphics[width=\columnwidth]{images/2_3.PNG}
177: The StockCode variable should ideally contain letters. So we have
filtered out the codes with only letters. We can observe from
Figure \ref{2.4}, different type of transactions based on these
(example D is for discounted transaction).\\
179: \begin{figure}
181: \label{2.4}
183: \includegraphics[width=\columnwidth]{images/2_4.PNG}
188: We have added a new variable to indicate total price of the
purchase (by multiplying unit price of each product with quantity
purchased). Each transaction corresponds to the prices for a
single product. On grouping the records based on a single order,
we can see the complete price for that order as shown in Figure
\ref{2.5}.\\
190: \begin{figure}
192: \label{2.5}
194: \includegraphics[width=\columnwidth]{images/2_5.PNG}
197: We can visualize the orders distinguished on the basis of total
price of the basket. It can be shown as Figure \ref{2.6} using a
pie-chart.
199: \begin{figure}
201: \label{2.6}
203: \includegraphics[width=\columnwidth]{images/2_6.PNG}
211: Upon checking, we found that there are 1483 keywords present in
the description variable of the dataset. The most common keywords
can be determined based on the occurrences. The Figure \ref{3.1}
shows the top word occurrences.
213: \begin{figure}
215: \label{3.1}
217: \includegraphics[width=\columnwidth]{images/3_1.PNG}
236: The Figure \ref{3.2} shows silhouette score for different values
of k. These scores do not have significant differences, but since
for k value greater than 5, the resulting clusters have very few

```

elements in them, we have taken k as 5.  
 238: \begin{figure}[H]  
 240: \label{3.2}  
 242: \includegraphics[width=\columnwidth]{images/3\_2.PNG}  
 249: From the silhouette plot shown in Figure \ref{3.3} we can see  
 that cluster 1 has more number of elements than the other  
 clusters. But overall distribution of elements in the clusters is  
 comparative. Same can be seen from the Figure \ref{3.4}.  
 251: \begin{figure}  
 253: \label{3.3}  
 255: \includegraphics[width=\columnwidth]{images/3\_3.PNG}  
 258: \begin{figure}  
 260: \label{3.4}  
 262: \includegraphics[width=\columnwidth]{images/3\_4.PNG}  
 267: The main idea of principal component analysis (PCA) is to reduce  
 the dimensionality of a data set consisting of many variables  
 correlated with each other, either heavily or lightly, while  
 retaining the variation present in the dataset, up to the maximum  
 extent. The initial matrix has large number of variables and  
 hence, PCA is used for dimensionality reduction. From the  
 Figure\ref{3.5} we can say that we need more than 100 components  
 to explain 90\% of the variance in the data.\\  
 269: \begin{figure}  
 271: \label{3.5}  
 273: \includegraphics[width=\columnwidth]{images/3\_5.PNG}  
 276: Another application of PCA is that it sets the indication of  
 cluster membership. Biplot is the best example that can be  
 provided here to support this idea. Using biplot, we get the  
 indication of number of clusters in a dataset. Below Figure  
 \ref{3.6} shows these on limited number of components (since it  
 is only for visualizing cluster distribution). We can observe the  
 groupings of points or clusters as expected.\\  
 278: \begin{figure}  
 280: \label{3.6}  
 282: \includegraphics[width=\columnwidth]{images/3\_6.PNG}  
 288: In the previous section, we have divided products in 5 clusters.  
 We have added a dummy variable categ\\_product to indicate the  
 cluster to which that customer belongs. Based on the clustering  
 done on products we have created variables categ\\_0..4 which  
 stores amount spent on each of the product category. And the  
 categ\\_product variable which we have just created will have  
 initial cluster assignment based on these variables. These can be  
 further grouped on the basis of InvoiceNo as shown in Figure  
 \ref{4.1}.\\  
 290: \begin{figure}  
 292: \label{4.1}

```

294: \includegraphics[width=\columnwidth]{images/4_1.PNG}
303: In the previous section we have seen the basket price of each
     invoices. For further analysis we will combine these on the basis
     of customerID to analyze the number of purchases made by each
     customer as shown in Figure \ref{4.2}. A customer category of
     particular interest is that of customers who make only one
     purchase. So one objective may be, for example, to target these
     customers in order to retain them. In the dataset we have almost
     one-third of the customer base similar to this.
305: \begin{figure}
307: \label{4.2}
309: \includegraphics[width=\columnwidth]{images/4_2.PNG}
317: Using the silhouette score, the optimum value of k comes out to
     be 11. The assignment of customers into different clusters is
     shown in Figure \ref{4.3}
319: \begin{figure}
321: \label{4.3}
323: \includegraphics[width=\columnwidth]{images/4_3.PNG}
330: There is a certain disparity in the sizes of different groups
     that have been created. So we have validated it using PCA. From
     the representation shown in Figure \ref{4.4}, it can be seen, for
     example, that the first principal component allow to separate the
     tiniest clusters from the rest. More generally, we see that there
     is always a representation in which two clusters will appear to
     be distinct.\\
333: \begin{figure}
335: \label{4.4}
337: \includegraphics[width=\columnwidth]{images/4_4.PNG}
342: As with product categories, another way to look at the quality of
     the separation is to look at silhouette scores shown in Figure
     \ref{4.5} within different clusters:\\
344: \begin{figure}
346: \label{4.5}
348: \includegraphics[width=\columnwidth]{images/4_5.PNG}
374: In the Python code, we have imported the module ‘linear\_model’
     from the ‘sklearn’ package to perform Logistic Regression by
     using the function ‘logistic\_regression’. And we have taken the
     $k=5$ for k-fold cross validation. While performing Logistic
     Regression, we created an instance of the Class\_Fit class and
     then ran the model on training data and see how the predictions
     are made as compared to the real values. The learning curve graph
     is as shown in Figure \ref{5.1}.
376: \begin{figure}
378: \label{5.1}
380: \includegraphics[width=\columnwidth]{images/5_1.png}
383: As we can see from the Figure \ref{5.1}, when the number of

```

training examples increases, the cross-validation and train curves almost converge towards the same limit suggesting that the model has low variance. Thus we can say that model is not suffering from over-fitting. Also one point to note is that the accuracy is high, which means that the model has low bias, thus suggesting that it does not under-fit the data. The precision which we got from running the Logistic Regression model on the training data is 88.78\%.

394: In Python, the ‘neighbors’ library is imported from the sklearn package which performs the KNN classification through the KNeighborsClassifier function. The parameters that are used are ‘n\\_neighbors’ which represents the number of neighbors to use, in our case we have used the np.arange method to give sequence from 1 to 49. Also, we run the model using the K-fold Cross Validation with the value of \$k=5\$. Once the model is run, we have drawn the learning curve graph which is as represented in the Figure \ref{5.2}.

396: \begin{figure}

398: \label{5.2}

400: \includegraphics[width=\columnwidth]{images/5\_2.png}

410: In Python, the ‘ensemble’ library is imported from the sklearn package which performs the Random Forest classification through the RandomForestClassifier function. The parameters given to this function are criterion, n\\_estimators and max\\_features. The criterion is used to measure the quality of the split. The Gini is for measuring the Gini impurity and Entropy is for information gain. The max\\_features are the number of the features that can be chosen when looking for the best split. For ‘sqrt’, the number of maximum features chosen are square root of the number of the features and for ‘log’, it is log of the number of the features. And the n\\_estimators is the number of trees in the forest. Once the model is run, we have drawn the learning curve graph which is as represented in the Figure \ref{5.3} .

412: \begin{figure}

414: \label{5.3}

416: \includegraphics[width=\columnwidth]{images/5\_3.png}

436: In Python, the ‘ensemble’ library is imported from the sklearn package which performs the Gradient Boosting classification through the GradientBoostingClassifier function. The parameter given to this function is n\\_estimators which is the number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number results in better performance. Learning curve is shown in the Figure \ref{5.4}.

438: \begin{figure}

440: \label{5.4}

442: \includegraphics[width=\columnwidth]{images/5\_4.png}

452: So now, we have the data available for two months, and through that we can define the category to which the consumer belongs. The predictions now obtained by running the classifiers on test data can be tested against these categories. The instance of the k-means clustering method that we used in the Customer Categories section is used to define the category to which a client belongs. This contains the predict method which will calculate the distance of the consumers from the centroids of the 11 categories that we deduced, and the category which is closest to the clients' buying habits will define his/her category. Thus all we need after this for the execution of the classifier is to select the variables on which it acts, i.e. on mean, cat\\_0, cat\\_1, cat\\_2, cat\\_3 and cat\\_4. After examining the predictions of the different classifiers, we get precision scores as shown in Table \ref{t:precisionscore}.\\"

454: \begin{table}[htb]  
457: \label{t:precisionscore}

figures 23

tables 1

includegraphics 23

labels 24

refs 25

floats 24

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

---

WARNING: algorithm and below may be used improperly

355: In the previous section, we have made different client categories. In this part we will adjust a classifier so that the consumers can be classified in different client categories. The main aim of this is to enable the Classification on the first visit of the customer. To do this, we have defined a class that will allow interfacing the common functionalities to the different classifiers. Since we are going to classify the client on the basis of his/her first visit, the only parameters that we take into consideration are the contents of the basket and not the frequency of visits or the variation in the basket price over a period of time. Once this is done, we have split the dataset into train and test sets. The classification algorithms which we used to do this are mentioned 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
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
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
```

# Big Data in Safe Driver Prediction

Jiaan Wang

Indiana University Bloomington  
3209 E 10 St  
Bloomington, IN 47408  
jervwang@indiana.edu

Dhawal Chaturvedi

Indiana University Bloomington  
2679 E 7th St  
Bloomington, Indiana 47408  
dhchat@iu.edu

## ABSTRACT

For years, people have been trying to reduce their automobile insurance bills. Insurance companies claim that price will be reduced for good drivers and raised for bad ones. However, inaccuracies in their data predictions lead to the exact opposite. The data-set being used is released by Porto Seguro, an auto and homeowner insurance company from Brazil. It consists of information from several hundred thousands of policyholders. The goal is to predict the probability an auto insurance policyholder files a claim the next year using classification algorithms. A good prediction with decent accuracy can correctly adjust prices for policyholders.

## KEYWORDS

i523, HID233, HID204, Big data, Classification, Safe Driving, Predictive Analytics, Neural Networks

## 1 INTRODUCTION

Everyday, people die from car accidents and it should come as no surprise that automobile accidents are one of the most common causes of death in the United States [8]. As reported by the CDC, Centers for Disease Control, approximately over 40,000 people lose their lives to fatal automobile accidents each year. It should be clear that we need to enhance road safety for drivers all over the states [5]. However, as we are currently in the age of big data, these automobile accidents could be prevented by using modern technologies and methods such as artificial intelligence and predictive analytics.

Big data describes large quantities of data that are impossible to analyze using traditional data analysis methods. It includes structured and unstructured data. Structured data can be SQL database stores and unstructured data can be videos, images, social media feeds, etc. In industries, data analytics is often performed on big data in order to find specific patterns or anomalies that could prove useful for business decisions and choices. The amount of data is usually irrelevant in these cases. For example, smart cars utilize big data to improve their safety features and systems. They collect data such as driving patterns and routes as they travel from point A to B. This information is then sent to the computers onboard and gets transferred to the company servers where the data undergo analysis. The result is then collected and stored to enhance smart-car systems [10].

Big data is also helpful in providing insights for product development. It can find the causes for issues and problems in products through data analytics, which can then be used to improve the design. For example, the driver assistance feature on a Mercedes-Benz car not only has safety features but also collects data on driving habits. If large amount of drivers speed through intersections or

break hard during traffic hours, the company can obtain these information and use them properly to enhance their systems for better road safety. They could add a detector with GPS data to spot intersections or traffic jams. Furthermore, another data that are useful to collect are driving routes. Google Street View utilizes big data from driving routes to update their maps and display views of different places [10].

Aside from smart cars, we also need to master data collection in order to know how automobile accidents happen. For example, the technology behind the famous black box, which tracks planes and cockpit communications to determine the reason behind crashes, is getting used in cars as well. It is not expensive nor complicated to apply this technology onto the majority of vehicles out there. By recording the precise time, locations, speed and other variables, this technology can definitely help us collect valuable data and information in car accidents. The result from data analysis can also assist us in a deep understanding of causes behind automobile collisions in order to save more lives by preventing future accidents. The first country who thought to implement this technology on cars was South Korea. As a result, in the following year, a 14 percent decrease was saw in the number of car accidents along with a 20 percent decrease in the number of injuries and deaths of fatal automobile accidents [5].

Predictive analytics is a powerful method in big data analytics to help predict future events or outcomes based on current and historical data. It usually utilizes big data techniques such as data mining, predictive modeling and statistics. It uses a wide range of predictive models which depends on the type of event we are predicting. For example, most predictive models produce a number called a score where a higher score indicates a higher chance of that outcome happening in the future. It is a very useful tool for making business decisions and assessing potential risks in many industries such as insurance, retail, etc. Predictive analytics does not inform users about things that has happened before today. It tries to predict for a particular driver the probability that he or she may be involved in an car accident in the near future or any other chose time as accurately as possible [8].

For example, in order to find high-risk drivers, it is not enough to just have driving records, automobile incident reports or traffic tickets. We also need something called telematics. Telematics is defined as the combination of telecommunications and informatics. It collects, stores, sends and receives data and information via transmission-enabled devices. For example, the use of the car black box technology mentioned previously to collect and obtain information on driving behaviors or patterns is called vehicle telematics. With telematics data, companies can determine the possibility of a driver in a future car accident along with the expenses coupled with it. They can also take actions such as putting high-risk drivers

into training schools to correct their bad driving behaviors before an incident happens [8].

By studying the telematics data from a specific driver, we can learn his or her driving behaviors and create a report that details the potential danger this driver may inflict and use these data to correct those bad driving habits. The probability of a driver being involved in a car accident in the future can be used to categorize drivers into different groups. With these probabilities, companies can create a safety score to provide them with suggestions on which drivers to deal with first. The fundamental role for a safety score is to identify drivers with high risks before an accident happen to give the driver a chance to correct those bad driving habits and prevent incidents from happening [8].

However, just like other countless programs on risk evaluation, predictive analytics is not supposed to be flawless. In companies such as UPS, FedEx, USPS or any other services that use a large amount of drivers and vehicles called fleet vehicles, predictive analytics is just the first process and it requires the total cooperation and commitment from the company, drivers and fleet staff to achieve the highest efficiency. Companies that have numerous successful fleet operations are those who plan ahead and bring together all fleet personnel into the action. The most effective way to adjust the accuracy and precision of predictive models is to test the models every few days or few weeks or any other time that is suitable for the operations. Due to the fact that most operations have tight schedules, this leeway time will give companies enough room to call in safety personnel to step in to either train drivers to correct their driving habits or repair fleet vehicles beforehand to avoid major system failures [8].

Predictive analytics and telematics data are being used in almost all fleet companies as more of them start to see the value in predictive analytics. With the help of predictive analytics, fleet companies can be actively engaged in making better decisions about their fleets and companies by improving road safety, reducing expenses and risks or decreasing work load time [8].

## 2 BIG DATA IN AUTO INSURANCE

For a long time, auto insurance companies have calculated insurance rates based on personal mileage through out the years [4]. Traditional auto insurance companies categorize users by demographics such as gender and other factors such as education. They then make predictions based on past statistics about their chances of getting involved in future car accidents. This means that the monthly payments insurance companies charge you are only calculated according to the information they have on you and these information has nothing to do with your driving behaviours. As a result, the premiums you pay every month is usually based on past data from people who have identical demographics as you. While a few of the factors are actually helpful in determining your risk score - for example, if you have had multiple accidents in the past, you are likely to be involved in a new one in the future - other factors such as how many cars you previously owned have little to do with your actual risk of being in an accident but yet they still matter when calculating the price. And no one ever use the most important factor in determining monthly premiums which is driving behaviors [6].

Major auto insurance companies have connection to large amount of information as well as data processing power in order to calculate risk scores and monthly premiums. It is no easy task for them to combine several different factors into one single price. Still, big data is not fully utilized. Even though these companies use a variety of different models to calculate monthly rates for drivers, not all of their methods are optimized. In a study conducted earlier this year, it was found, with the help of data mining, that there exists predictive models that have higher accuracy in categorizing drivers into high and low risk groups. Among those models was one that combined 16 factors which produced an extremely high accuracy in risk assessment [6].

However, powered with advanced technology and big data analytics, insurance companies now-days have access to customers' driving behaviours for more personalized insurance rates. They collect specific data on how often you drive every day, how long you drive each time, how often you speed, how often you break hard and so on to determine the probability of you being in an accident in the near future. With these precise data on each individual customer, insurance companies can assess risk scores for everyone and use that information to calculate your monthly rates [4]. For example, an auto insurance company called Root is one of the first mobile auto insurance companies that are intended to help you on the go. They promise to only insure the good drivers to make sure they get the best rates. Their methods are simple. Download the app, take a test drive with the app on-board for several weeks and Root will send a personalized premium plan based on your driving behavior. Then you can just select the plan you want to purchase and buy via your phone [6].

These new and innovative mobile auto insurance plans are called UBI, short for Usage-Based Insurance and they calculate monthly premiums mostly based on driving habits. Applications on smart-phones and on-board diagnostic devices along with in-car tracking technology from manufacturer are used to record mileage and driving behaviours. These mobile insurance programs tend to give discounts for good drivers as a reward for their good driving behaviors. They are even adding new rewards such as roadside assistance on top of discounts for drivers who have maintained good driving behaviors for long periods of time. By collecting big data on driving habits, auto insurance companies can promptly discover mistakes when accidents happen by knowing the exact positions of each car and driving habits data such as speeding or braking as well as environmental data such as weather or road conditions [7].

On the surface, these Usage-Based Insurance plans appear to be plausible and feasible. An application or a sensor is installed on your phone or car to track your driving behavior instead of estimating costs based on factors such as age, gender, education, traffic records, accident reports and so on. Several programs such as *Drivewise* from Allstate insurance and *Snapshot* from Progressive insurance have been released to the public for a couple of years in some states, completely based on customers' choices. You do not have to install them if you do not want to. However, the majority of drivers have been embracing these monitoring devices since it has no apparent downside. As long as your driving behaviors are considered to be safe, such as slow braking and accelerating or no driving around midnight, your should receive discounts like 5 to 10 percent or even up to 20 percent on your monthly premiums. On the other hand,

customers are starting to worry about their privacy but we still do not know what the worst thing that might happen if we keep letting insurance companies monitor our driving behavior. According to the Wall Street Journal, these Usage-Based Insurance plans are growing exponentially. The biggest auto insurance company in United States, State Farm, announced their plans to expand their *Drive Safe and Save* program to the entire country soon. Their major advertising strategy is that by enrolling in the program, consumers can get discounts in their insurance premium by proving that they drive safely [9].

For example, one of the earliest Usage-Based Insurance programs available to the public was released about 10 years ago by Progressive and General Motors. This particular program, with the help of GPS, applied discounts based on customer mileage. Many of the Usage-Based Insurance programs these days still implement this strategy but many improvements have been made. Insurance companies nowadays know everything about the way you drive from where you drive to when you drive as well as how you drive. There are also a variety of options to choose from such as *pay as you drive* and *pay how you drive*, thanks to telematics. The advantage of having telematics in these insurance programs is to improve efficiency such as reducing response time for accidents. In addition, Usage-Based Insurance data can be analyzed using on-board diagnostic devices which are often plugged in via the on-board diagnostic 2 port on cars. These diagnostic devices do not have the ability to track car positions but they do generate more precise and meticulous data about car usage. Although telematics is the typical way to record driving habits, new creations in the future will possibly use smart-phone's location services or GPS abilities to track bad driving behaviors such as speeding and hard braking. Liberty Mutual and State Farm both tested their new tracking technology via smart-phones or other smart devices on-board cars in 2015. By 2020, the majority of auto insurance companies will be using Usage-based Insurance programs coupled with telematics data. It is no doubt that Usage-Based Insurance programs will continue to grow and achieve even higher precision and availability [3].

### 3 CURRENT APPLICATIONS

Predictive analytics are being utilized with telematics data to enhance and improve road safety. Telematics have been used for a long time in insurance industry to monitor driving behaviors such as speeding to identify high-risk drivers. Now coupled with predictive analytics, these data are being analyzed to predict the likelihood of a driver being involved in future accidents. SmartDrive Systems, a transportation safety and intelligence company, employs even more interesting ways to predict accidents. They record and gather video feeds from dashboard cameras in cars which are then integrated with telematics data. This way, they can improve their predictive analytics on driver safety and eventually leads to better predictions [1].

SmartDrive Systems uses a private cloud to provide their clients with predictive analytics solutions. All the data from their clients are collected by the company and stored on their cloud. The data includes telematics and video feeds from millions of clients with more than 4 billion mileage. SmartDrive constantly improves their predictive models because the agreements SmartDrive has with

their clients permit them to study all the data they gather. The usage of both telematics data and video feeds is a great idea which enables SmartDrive researchers to better understand the data and interpret the results. By combining what they see through the video feeds and the results from telematics data analysis, the researchers can draw conclusions such as making a U-turn on a narrow road within some fixed radius is dangerous [1].

Indiana State Police came up with a different way to predict incidents and are making their predictive analytics methods open source. In their approach, they produced something called *Daily Crash Prediction Map* which finally completed in November. It contained data such as accident reports from all the police departments in Indiana going back to 2004 as well as data on daily weather, historical traffic amount and so on. This map highlights where potential accidents may happen categorized by their probabilities. It also features information about past accidents such as locations, dates, causes, fatalities and so on [2].

Liberty Mutual is the nation's third largest property and casualty insurance company. Last year, Liberty Mutual partnered up with Subaru. In doing so, customers was granted access to *Starlink*, Subaru's multimedia and navigation system, which can track and notify drivers via an app if they are speeding or braking too hard. By enrolling in the *RightTrack* program provided by Liberty Mutual, drivers can get up to 30 percent discounts for good driving behaviors [4].

## 4 DATA ANALYSIS

The data we are gonna use for our analysis is of Porto Seguro. It is one of Brazil's largest auto and homeowner insurance companies. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. The task is to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year or not. An accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

### 4.1 Approach

We will be mainly discussing about the Exploratory data Analysis we have performed on the data until now. We will be using the help of both R and python environment and supporting packages to perform the necessary statistical analysis. Along with this, we will discuss about the Machine or Deep Learning algorithms / models that we will be using to achieve the near solution for the problem.

### 4.2 Feature Information

Dimensions of the data [Rows x Features] : [ 595212, 59 ]

The data-set constitutes different varieties of features.

**Binomial Features** [Count : 17] : ps.ind\_06.bin, ps.ind\_07.bin, ps.ind\_08.bin, ps.ind\_09.bin, ps.ind\_10.bin, ps.ind\_11.bin, ps.ind\_12.bin, ps.ind\_13.bin, ps.ind\_16.bin, ps.ind\_17.bin, ps.ind\_18.bin, ps.calc\_15.bin, ps.calc\_16.bin, ps.calc\_17.bin, ps.calc\_18.bin, ps.calc\_19.bin, ps.calc\_20.bin.

**Categorical Features** [Count : 14] : ps.ind\_02.cat, ps.ind\_04.cat, ps.ind\_05.cat, ps.ind\_01.cat, ps.ind\_02.cat, ps.ind\_03.cat, ps.ind\_04.cat, ps.ind\_05.cat, ps.ind\_07.cat, ps.ind\_06.cat, ps.ind\_08.cat, ps.ind\_09.cat, ps.ind\_10.cat, ps.ind\_11.cat.

**Integer Features** [Count : 16] : ps.ind.01, ps.ind.03, ps.ind.14, ps.ind.15, ps.ind.11, ps.calc.04, ps.calc.05, ps.calc.06, ps.calc.07, ps.calc.08, ps.calc.09, ps.calc.10, ps.calc.11, ps.calc.12, ps.calc.13, ps.calc.14.

**Floating Features** [Count :10] : ps.reg.01, ps.reg.02, ps.reg.03, ps.calc.01, ps.calc.02, ps.calc.03, ps.car.12, ps.car.13, ps.car.14, ps.car.15.

The remaining two features constitutes **id** and the **output (or target)**. All the features has been clearly represented using post script, **.cat** for categorical data, **.bin** for binomial data.

The **missing values** in the features are represented by -1.

### 4.3 Data Pre-processing

As shown in 1, missing values are found in 14 of the 58 columns. There are 6 features with more than 5000 missing row values. Owing to the shear size of the unavailable data, we have not performed any missing value treatment and removed these features from consideration. Of the remaining data, across rows, data is unavailable in almost 500 (<1%) rows and these are promptly removed.

[Figure 1 about here.]

### 4.4 Distribution of the target variable

As shown in 2, target variable claims is a binary variable with a skewed distribution of classes. 96% of the customers didn't make any claims. We wish to consider this distribution in measuring classification accuracy. Area under the ROC curve, recall and precision would be relevant metrics in this case.

[Figure 2 about here.]

### 4.5 Numerical Predictors (vs) Target Variable

Average value of most of the numerical predictors is higher when the claims are filed. This is a unique phenomenon and we intend to use this contrast in predicting the target variable.

### 4.6 Artificial Neural Networks

Artificial neural networks (ANNs) are computing models which are based on biological neural networks that constitute human brains. The idea of ANNs is based on the belief that working of human brain can be imitated for computers by using silicon and wires as living neurons. Such systems learn by progressively improving their performance to do tasks by considering examples, generally without task-specific programming. The human brain can be considered as a complex network of nerve cells called neurons(about 86 billion). They are inter-connected to other millions of cells by Axons. These neurons then react to stimulation from external environment or inputs from other organs. A neuron can then send the message to other neuron to handle the issue or does not send it forward.

ANNs also try to imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation. Each node is assigned with weight. ANNs are capable of learning, which takes place by altering weight values. If the network generates the desired output, there is no need to adjust the weights. However, if the network

generates an undesired output or an error, then the system alters the weights in order to improve subsequent results.

### 4.7 Types of ANNs

**4.7.1 Feedback ANN.** In these type of ANN, the output goes back into the network to achieve the best-evolved results internally. The feedback network feeds information back into itself and is well suited to solve optimization problems. Feedback ANNs are used by the Internal system error corrections.

**4.7.2 Feed Forward ANN.** A feed-forward network is a simple neural network consisting of an input layer, an output layer and one or more layers of neurons. Through evaluation of its output by reviewing its input, the power of the network can be noticed base on group behavior of the connected neurons and the output is decided. The main advantage of this network is that it learns to evaluate and recognize input patterns.

**4.7.3 Radial Basis Function Neural Network.** The RBF neural network is the first choice when interpolating in a multidimensional space. The RBF neural network is a highly intuitive neural network. Each neuron in the RBF neural network stores an example from the training set as a fiprototypef. Linearity involved in the functioning of this neural network offers RBF the advantage of not suffering from local minima.

**4.7.4 Kohonen Self-Organizing Neural Network.** Invented by Teuvo Kohonen, the self-organizing neural network is ideal for the visualization of low-dimensional views of high-dimensional data. The self-organizing neural network is different from other neural networks and applies competitive learning to a set of input data, as opposed to error-correction learning applied by other neural networks. The Kohonen self-organizing neural network is known for performing functions on unlabeled data to describe hidden structures in it.

**4.7.5 Recurrent Neural Network.** The recurrent neural network, unlike the feed forward neural network, is a neural network that allows for a bi-directional flow of data. The network between the connected units forms a directed cycle. Such a network allows for dynamic temporal behavior to be exhibited. The recurrent neural network is capable of using its internal memory to process arbitrary sequence of inputs. This neural network is a popular choice for tasks such as handwriting and speech recognition.

**4.7.6 Classification-Prediction ANN.** It is the subset of feed-forward ANN and the classification-prediction ANN is applied to data-mining scenarios. The network is trained to identify particular patterns and classify them into specific groups and then further classify them into finovel patternsf which are new to the network.

**4.7.7 Physical Neural Network.** This neural network aims to emphasize the reliance on physical hardware as opposed to software alone when simulating a neural network. An electrically adjustable resistance material is used for emulating the function of a neural synapse. While the physical hardware emulates the neurons, the software emulates the neural network.

## 4.8 Data Analysis Using Neural Networks

Rather than beginning our inquiry into the data-set with more traditional methods like regression we straight away tried to learn Artificial Neural Networks. Logistic regression itself, can be thought of as a special case of a neural network with a single neuron (perceptron).

After studying and learning the theory behind ANNs we proceeded to learn how to implement them – by ourselves at first, and later using TensorFlow. So far we have tried quite a bit of different models and learned some lessons about the data.

- **Data Cleaning :** As pointed out by the EDA above, there were a few columns which had a lot of missing data. For columns which had > 1000 values missing (6 columns), we disregarded them altogether. For the remaining data points, we disregarded the rows which had any one particular value missing. We started with a simple data cleaning strategy so as to not complicate it too much at the initial stages, but we will probably want to look at it again as we go along.
- The first thing that we tried is using a simple **perceptron**. The input layer had 51 nodes (after removing the id, target and 6 other columns in data cleaning) and the output layer had a single perceptron with a sigmoid activation function. The best score that we got when we uploaded our code to Kaggle was 0.03 whereas the leader-board is hovering around 0.290 so this is not too impressive.
- However, now we added more hidden layers and nodes to see if we get a better job of fitting the data. To start off, we only consider the continuous variables so that we don't have to worry about handling binary/categorical data. We have 24 nodes in the input layer. The final layer has one node since it is a classification problem. We kept all activation to be logistic and experimented a bit with the number of hidden layers and nodes to get a best score of **0.211** with this simple approach.
- **Issued Encountered** Looking at the results from the neural network there is one major issue. Whenever we add too many hidden layers (> 3) the outputs for the test data are all  $\approx 0$  and the score drops. After some trial and error, we have diagnosed the issue to be the biased nature of the training data (96% of the training data are 0's). So the ANN sees too many zeros and consequently predicts mostly zeros. We aim to explore different sampling methods to train the neural network to improve the results further (along with cross validation which we haven't implemented).

## 5 OTHER TECHNIQUES THAT CAN BE USED

Among the machine Learning algorithms that are used in practice, gradient tree boosting is one technique that shines in many applications. Tree Boosting has been shown to give many state of the art results for many standard classification problem.

The most important factor for the success of XGBoost is its ability to scale in all the scenarios. The XGBoost algorithm runs ten times faster than the existing popular solutions on a single solution and scales to billions of examples in distributed or memory-limited settings.

The Porto Seguro data-set is clearly an classification problem, the data will be having only two outputs either the car insurance holder is going to claim or not.

While domain dependent analysis and feature engineering plays an important role in defining or modeling the solutions, the fact that XG Boost is the consensus choice for learners shows the impact and importance of our system in tree boosting. One problem with the Porto Seguro data-set we do not have have much information about the Features and all the feature must have to undergo through strict statistical treatment to build an optimal solution.

## 6 CONCLUSION

Reducing insurance rates has always been a difficult task in the past. However, armed with advanced technology, insurance companies now have the ability to track personal driving behaviours to provide better suited personalized insurance plans. We performed Neural Networks algorithm on clients data from Porto Seguro, one of Brazil's largest auto insurance companies in order to predict the probability of drivers filing claims in the next year. Our analysis proved to be a success and our model yielded a high accuracy. In the future, we can use other machine learning algorithm or classification techniques to compare the accuracy of different models.

## 7 APPENDIX

### 7.1 Links to iPython notebook:

[https://github.com/bigdata-i523/hid233/blob/master/project/Shallow-Neural\\_Nets.ipynb](https://github.com/bigdata-i523/hid233/blob/master/project/Shallow-Neural_Nets.ipynb)

### 7.2 Links to iPython notebook pdf version:

[https://github.com/bigdata-i523/hid233/blob/master/project/Shallow-Neural\\_Nets.pdf](https://github.com/bigdata-i523/hid233/blob/master/project/Shallow-Neural_Nets.pdf)

### 7.3 Work Contribution

Jian Wang - Sections: Abstract, Introduction, Big data in auto insurance, Current application, Data analysis (10 percent) and Conclusion

Dhawal Chaturvedi - Sections: Data analysis (90 percent), Other techniques that can be used and Conclusion as well as Jupyter notebook

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] Steve Banker. 2016. Using Big Data And Predictive Analytics To Predict Which Truck Drivers Will Have An Accident. Web Page. (Oct. 2016). <https://www.forbes.com/sites/stevebanker/2016/10/18/using-big-data-and-predictive-analytics-to-predict-which-truck-drivers-will-have-an-accident/#fd6888b1cb0> HID: 233, Accessed: 2017-11-28.
- [2] Jenni Bergal. 2017. Troopers Use Big Data To Predict Crash Sites. Web Page. (Feb. 2017). [https://www.huffingtonpost.com/entry/troopers-use-big-data-to-predict-crash-sites\\_us\\_589c88ebe4b0985224db5e19](https://www.huffingtonpost.com/entry/troopers-use-big-data-to-predict-crash-sites_us_589c88ebe4b0985224db5e19) HID: 233, Accessed: 2017-11-28.
- [3] Crosley Law Firm. 2016. Benefits and Concerns About Usage-Based Insurance. Web Page. (Nov. 2016). <https://crosleylaw.com/blog/big-data-behind-bad-driving-insurers-use/> HID: 233, Accessed: 2017-11-28.
- [4] Brian Fung. 2016. The big data of bad driving, and how insurers plan to track your every turn. Web Page. (Jan. 2016).

- <https://www.washingtonpost.com/news/the-switch/wp/2016/01/04/the-big-data-of-bad-driving-and-how-insurers-plan-to-track-your-every-turn/> HID: 233, Accessed: 2017-11-28.
- [5] Mikkie Mills. 2017. 4 Ways How Big Data Will Improve Road Safety. Web Page. (May 2017). <https://datafloq.com/read/4-ways-big-data-will-improve-road-safety/3127> HID: 233, Accessed: 2017-11-28.
- [6] Cristol Rippe. 2017. Big Data, Better Rates: Why Current Car Insurance Rate Calculations are Unfair. Web Page. (Jan. 2017). <https://blog.joinroot.com/big-data-better-rates-why-current-car-insurance-rate-calculations-are-unfair/> HID: 233, Accessed: 2017-11-28.
- [7] Jonathan Shafer. 2016. How Big Data Analytics is Changing the Competitive Auto Industry. Web Page. (Aug. 2016). <https://pentaho.com/blog/2016/08/26/how-big-data-analytics-changing-competitive-auto-industry> HID: 233, Accessed: 2017-11-28.
- [8] Grace Suizo. 2015. Using Predictive Analytics to Improve Fleet Decisions. Web Page. (Sept. 2015). <http://www.automotive-fleet.com/channel/gps-telematics/article/story/2015/10/using-predictive-analytics.aspx> HID: 233, Accessed: 2017-11-28.
- [9] Brad Tuttle. 2013. Big Data Is My Copilot: Auto Insurers Push Devices That Track Driving Habits. Web Page. (Aug. 2013). <http://business.time.com/2013/08/06/big-data-is-my-copilot-auto-insurers-push-devices-that-track-driving-habits/> HID: 233, Accessed: 2017-11-28.
- [10] David Walker. 2017. How Will Big Data From Self-Driving Cars Influence Road Safety. Web Page. (July 2017). <https://technofaq.org/posts/2017/07/how-will-big-data-from-self-driving-cars-influence-road-safety/> HID: 233, Accessed: 2017-11-28.

#### LIST OF FIGURES

1	Missing Data	8
2	Target Variable Distribution	9

Features	Missing Values Count
ps_ind_02_cat	216
ps_ind_04_cat	83
ps_ind_05_cat	5809
ps_reg_03	107772
ps_car_01_cat	107
ps_car_02_cat	5
ps_car_03_cat	411231
ps_car_05_cat	266551
ps_car_07_cat	11489
ps_car_09_cat	569
ps_car_11	5
ps_car_12	1
ps_car_14	42620

Figure 1: Missing Data

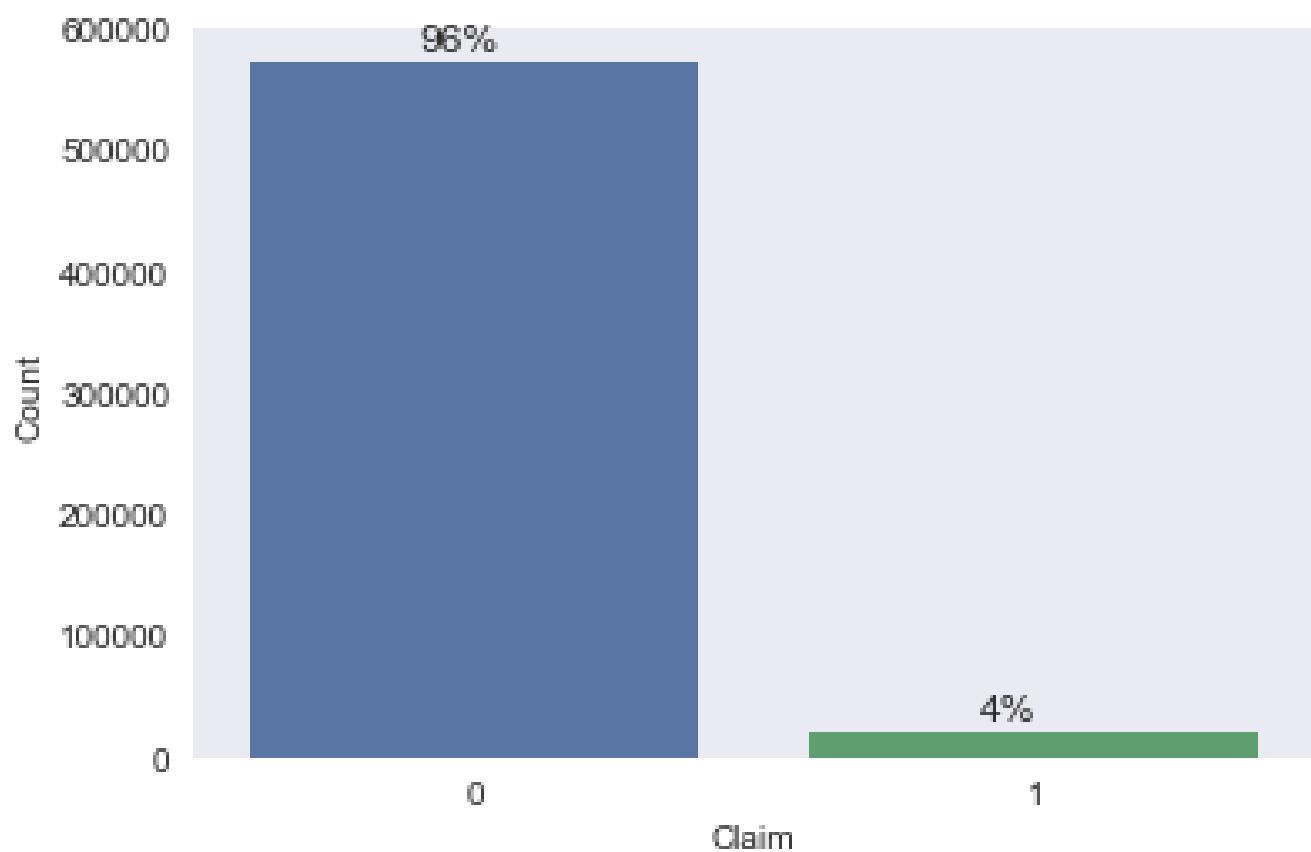


Figure 2: Target Variable Distribution

## 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

[2017-12-10 13.49.37] pdflatex report.tex

```
36:66      error    trailing spaces  (trailing-spaces)
37:70      error    trailing spaces  (trailing-spaces)
38:69      error    trailing spaces  (trailing-spaces)
72:69      error    no new line character at the end of file  (new-line-at-end-of-file)
```

---

## Compliance Report

---

```
name: Wang, Jiaan
hid: 233
paper1: Nov 03 17 100%
paper2: Nov 10 17 100%
project: 10%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
9
wc 233 project 9 4869 content.tex
wc 233 project 9 5070 report.pdf
wc 233 project 9 564 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
find input{format/final}
-----
passed: False

floats
-----
116: As shown in \ref{f:missing}, missing values are found in 14 of
the 58 columns. There are 6 features with more than 5000 missing
row values. Owing to the shear size of the unavailable data, we
have not performed any missing value treatment and removed these
features from consideration. Of the remaining data, across rows,
data is unavailable in almost 500 ( $\approx 1\%$ ) rows and these are
promptly removed.
119: \centering\includegraphics[width=\columnwidth]{images/missingdata
}
120: \caption{Missing Data}\label{f:missing}
124: As shown in \ref{f:numerical}, target variable claims is a binary
variable with a skewed distribution of classes. 96\% of the
customers didn't make any claims. We wish to consider this
distribution in measuring classification accuracy. Area under the
ROC curve, recall and precision would be relevant metrics in this
case.
127: \centering\includegraphics[width=\columnwidth]{images/target}
128: \caption{Target Variable Distribution}\label{f:numerical}
```

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

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
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 8220
non ascii found 8221
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

# **Big Data Analytics and Applications in the Travel Industry and Its Potential in Improving Travel Accessibility**

Weixuan Wang

Indiana University Bloomington

Bloomington, Indiana 47405

wangweix@indiana.edu

## **ABSTRACT**

Big data applications and analytics have been influencing and improving tourists' experience. Travel accessibility refers to provide access for people with disabilities or limited mobility (such as seniors), who represent a growing market in the travel industry by spending billions on leisure and business trips. This report explored the implementation of big data analytics and applications in tourism, disabilities related studies and assistive technologies for people with disabilities. This report explored the potentials of big data applications and analytics in understanding the needs and travel experience of people with disabilities and improving travel accessibility and quality of life for people with disabilities.

## **KEYWORDS**

i523, HID234, Big Data Analytics, Travel Accessibility, People with Disabilities, Quality of Life

## **1 INTRODUCTION**

People with disabilities represented a large neglected tourism market. According to Amadeus annual report, 15 percent of worldwide population (around 1 billion people) lives with some forms of disability [3]. According to United Nation, people with disabilities are the largest minority group in the world [4, 16, 19]. Notably, the number of people with disabilities is expected to increase as a result of extension of human life-span, decreases in communicable diseases, the improvement of medical technology, and decrease of child mortality [42]. While some forms of disabilities might be genetic, but temporary or permanent disabilities can happen to anyone, such as spinal cord injury after car accident, or limited mobility at later stage of life [19].

Population aging trend also signifies that disability will be a more common and urgent issue in the future [22]. The World Health Organization estimates that by 2050, 21.5 per cent of the global population will be aged over 65 [3]. As a large and fast growing minority group worldwide, people with mobility limits and accessibility issues faces a large ranges of barrier when traveling, and travel and tourism demand of this group is often underestimated or completely ignored [3]. According to the Open Door Organization (ODO) market report in 2015, people with disabilities spend 17.3 billion dollars annually for their own travel [33]. Because people with disabilities usually needs a care giver or family member to accompany them when traveling, the potential economic impact could double [33].

Accessible travel or accessible tourism refers to the inclusive travel activities that enable people with access requirements, including mobility, vision, hearing and cognitive dimensions of access,

to function independently, with equity and dignity through the delivery of universally designed tourism products, services and environments [3]. However, the travel experiences for people with disabilities are more than access issues. In order to achieve travel accessibility, which means provide travel activities for people with disabilities, a variety of aspects for travel needs must be taken in consideration. An accessible destination and appropriate accommodation only lay the foundation for a particular travel experience to happen for people with disabilities [33]. More aspects that need to consider for people who are traveling with disabilities, such as accessible transportation, accessible online booking [3].

The ultimate aim for those involved in supporting accessible travel is to empower every individual to plan and travel independently, at their own will [50]. However, the task is not a easy one. Making the whole travel chain accessible, including the information and booking procedures, as well as the infrastructure and processes become a important task for travel accessibility [3].

The development of information communication technologies especially the creation and distribution of user-generated content (UGC) or consumer-generated content (CGC) has successfully changed how people travel and how people gather information for travel [12]. Big data application and analytics has become a trending topic for the tourism industry and tourism studies [12]. The fast development of information and digital technology has changed many people's lives, especially the life of people with disabilities has also been improved by technology [10]. People with poor visions can using cell phones to contact others, access information online with screen readers. People with hearing problems can text other people with their cell phone. The use of big data for disabilities related research, disability informatics and developing assistive technology has been studies to improve the quality of life for people with disabilities [22].

Although big data is becoming an important topic in both tourism studies and disability related studies. There are a gap in the literature about how big data can be used in accessible tourism practice and studies, the potential of big data analytics and applications for improving travel accessibility has not been discussed before. Travel for business and leisure, especially travel independently and with dignity, constitute an essential needs for people with disabilities, and plays a fundamental part in the quality of life for people with disabilities. This study is trying to explore the use of big data applications and big data analytics in tourism and disability related practice and research, illustrating and discovering the potential of using big data applications and analytics for accessible travel and tourism practice and studies.

## 2 TOURISM AND BIG DATA

Information Communication Technologies (ICTs) have been transforming tourism business globally and revolutionizing the world of Tourism. It transforms tourism from a labor-intensive to an information-intensive industry [45]. Tourists influence by the developments in search engines, network speed and capacity have been using use technologies for better planning and experiencing their trips [47]. In addition, ICTs enable travelers to access reliable and accurate information and make reservations faster, cheaper and more convenient than the traditional way [12]. The development of ICTs also enables Internet users to both create and distribute information (especially multimedia information), which is called user-generated content (UGC) or consumer-generated content (CGC) [12].

Big data is a new and trending topic in the tourism industry and tourism studies, however, it is not unfamiliar to tourist activities. Most activities in the tourism industry had been generating a huge amount of data for several years. Booking flight tickets, reserving a hotel room and renting a car all leaves a data trail [40]. These data could add up to more than hundred of terabytes or petabytes structured data in the conventional databases [2]. Discussions of travel planning on online travel community such as the Lonely Planet Community, status updates and posts on social media like Facebook and Twitter, compliments and compliant on review websites like TripAdvisor and Yelp, recording and sharing travel experience on travel blogs constructs more challenging and live unstructured data that arrives at a much faster pace than a conventional database [2]. Tourism practitioners and tourism scholars are trying to understand tourists' behavior by accepting and analyzing these big data [40].

Tourists in the digital age often use a variety of tools to access information that the tourism industry or other users have provided [46]. A tourist produces a high volume of data when they are searching for travel websites, reporting issues on mobile applications, sharing traffic information in the cities, searching and posting on social media, taking and sharing photos, reporting experience on travel websites and social media, documenting their trips on blogs [2, 40]. All these data that are produced constantly can demonstrate tourists' motivation, interests, and their planning patterns and so on [47].

Previous studies have demonstrated several different usage and formats of big data in the travel and tourism industry [47]. Social media is one of them that has a huge effect on the tourism industry. Social media includes social networks, review sites, blogs, media sharing, and wikis [46]. The exceptional growth of these data sources has inspired companies and institutions to come up with new strategies to understand the socio-economic phenomenon in various fields [40]. Discussions and information sharing on social media are considered as electronic word-of-mouth (eWOM) that has in some degree substituted tradition face-to-face word-of-mouth (WOM) for information exchange of tourist experience [12].

Most tourism research utilizing big data are focusing on CGC or UGC, especially online reviews for a hotel. A recent study conducted by Guo, Barnes and Jia used data mining approach and linguistic analysis to extract meaning from 266,544 online reviews for 25,670 hotels [24]. They mined their customer review data from

TripAdvisor using a web crawler [24]. Through their linguistic analysis of their data and cross-comparing with perceptual mapping of the hotels, they found 19 controllable dimensions that are important for hotels to manage their interactions with visitors (such as the price for value, check in and check out) [24].

Photo post on photographic sharing website also can also provide extensive information on the tourists. Previous studies have connected photos posted on Panoramio, Flickr, and Instagram [10, 29]. Because when a tourist post pictures on these websites, their photo is tagged with geographic locations and ordered chronologically. Therefore analyzing photos posted by tourists can provide a photo density map to better understand tourists' behaviors, and potentially provide opportunities to detect atypical tourists behavior and characterize communities behaviors [29]. However, the study also has its own limitation because of the limitation of technology to better exploit the data [10]. Another study focused on the sequence of locations in shared geotagged photos by tourist to identify and recommend travel routes which helped the travel recommender system to generate personalized recommendation according to interests and time available [29].

Overall for tourism industry and tourism research, big data has becoming more and more popular. Both tourism practitioners and tourism researcher has recognized the influence of big data and big data sources for tourism development. Big data in the tourism industry are generated by tourists directly, compared to traditional data sets that are gathered from surveys, they have argued that these direct data from tourists themselves can better represent their true travel experience [24]. Therefore, big data presented us opportunities to better understand tourist behavior, their motivations, and interests.

## 3 DISABILITY AND BIG DATA

There are many different definition of disabilities from different organizations. The most cited official definition is the 1976 definition of the World Health Organization [4]: "An impairment is any loss or abnormality of psychological, physiological or anatomical structure or function; a disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being; a handicap is a disadvantage for a given individual, resulting from an impairment or a disability, that prevents the fulfillment of a role that is considered normal (depending on age, gender and social and cultural factors) for that individual". While people with disabilities are those people who have limitations in their actions or activities resulting from physical, sensory or cognitive impairments, however, there are many types and levels of disabilities and their actions and activities are affected differently by their disabilities [4]. The complexity of disabilities presents difficulties and challenges to accommodate the different needs of people with disabilities and improve their qualities of life [35].

The number of individuals living with some sensory or cognitive impairment or assisting an affected person is enormous [38]. Researchers has been using disability informatics to better understand people with disabilities. Disability informatics is a sub-specialty of health informatics that is defined as "any application that collects, manages, and distributes information that are related to people

with disabilities, as well as to caregivers (including familiar members and health care providers) and rehabilitation professionals" [4]. Disability informatics is closely related to other health informatics areas such as medical informatics, public health informatics and consumer health informatics, because people with disabilities usually have some secondary medical condition such as poor health status and increased personal health care needs.

Gather medical and health information can help to better understand and accommodate people with disabilities [38]. A study from the early 2000 has identified the potential of public health informatics for prevention at all vulnerable points in the causal chains leading to disability and proposed that applications should not be restricted to particular social, behavioral, or environmental contexts, but in a more global context [48].

Another previous research has designed and deployed an extended version of Artemis system (a cloud system designed to acquire data and store physiological data of clinical information for real-time analytics) in a hospital. They have identified that high speed physiological data produced at intensive care units as big data, and the proper use of such data can promote health, reduce mortality and disability rates of critical condition patients and create new cloud-based health analytics [26]. Research also has shown that many disabilities are genetic, therefore, bioinformatics has implications in the education of genetic screening and gen therapy treatments in the future [4].

People with disabilities usually need some assistive technology in their daily life. These technology that assist them to perform basic physical and social functions. The use information technology and assistive applications in disability informatics are categorized into three areas: virtual, personal, physical.

- Virtual environment refers to use of digital technologies like website and the Internet [4]. The digital revolution had and will continue to have a profound positive impact on the life of people with disability by empowering them with the help of digital technologies [4]. However, there are still access issues in the digital world. One of the barrier is the use of the World Wide Web (WWW or Web). Therefore, virtual environment for people with disabilities is usually discussed regarding to web accessibility.
- Personal Environment refer to having a safe personal environment for people with disabilities, which includes personal management and health monitoring [4]. Safety monitoring and health monitor devices are essential in this personal environment, which enables a safer personal environment and also provide health information for their medical care providers [4]. However the ethic of such health monitor devices are always in debate, some believe it can be an invasion of privacy and a restriction of personal freedom, others hold the ground that its main purpose is to help people with disease or disabilities, since it can alert their caregiver if the individual are exposed to harm (such as a person with mental disability and has a history of self-harming, these device can prevent unwanted behavior [15].
- Physical environment refers to the actual living space, traveling environment for people with disabilities. People with

mobility disabilities, visual impairment or cognitive limitation all need special help in their physical environment. Since the American Disabilities Act passed in the 1990s, the accessibility of physical environment has been improved in a great degree. However, people with disabilities still would meet some barrier and problem, one of them is the lack of curb cuts. Assistive information technologies has been developed in an effort to solve this problem. One of them is MAGUS, which is a project using geographical information system to inform users about wheelchair accessibility in urban areas [4]

The contribution of Big data and cloud computing have been recognized and accepted by researchers in health informatics [44]. The potential of big data and cloud computing for disability informatics and for people with disabilities has been explored by a few researchers and organizations. Data-Pop Alliance is one of the organization has recognized the big data and potential for study and help people with disabilities for disability informatics and people with disabilities [35]. Their research has categorized three type of big data source used across disability research: exhaust data (mobile-based data, financial transaction, transportation and online trace), digital content (social media and crowded-sourced/online content), and sensing data (physical and remote) [35]. They also provided the potential for some of these data sources, for example, researchers can use transaction data to compare cost, availability, and use of services that offer accessible options (such as accessible hotel listings) [35]. They also suggested that researcher can use social media data to represent people with disabilities as a network of interaction and using crow-sourcing to map the locations of accessible businesses and public places [35]. The organization has also identify four functions of big data on disability: descriptive, predictive, diagnostic and engagement. Descriptive function of big data is to describing and presenting the collected information such as using location data to map workplaces that are accessible to people with disabilities [35]. Predictive function is making inferences based on collected information such as discovering trends in the growth of number of accessible businesses in a certain urban area, while the diagnostic function means establishing and making recommendations on the basis of causal relations such as showing what can help increasing accessible business in a certain area [35]. Finally, the engagement function refer to shaping dialogue within and between communities and with key stakeholders through communication of data [35].

Cloud computing in combined with big data can also provide great opportunities for research and improvement of quality of life for people with disabilities [7]. The term cloud "refers to everything a user may reach via the Internet, including services, storage, applications, and people" [25]. Depending on the type of using, the "cloud" can be used for different purpose, such as for companies, the cloud could be used for hosting services so as to avoid the costs and difficulties associated with hosting one's own servers and software and for individuals, the could is often used as information storage [26]. Regardless of the types of usages for cloud, the end using must still access the information and services residing in the cloud through device like a smart phone or computer [25]. Cloud computing has been used to provide more accessible virtual environment,

especially Web access through project like WebAnywhere, which is a cloud based tool for blind using to access Internet [25].

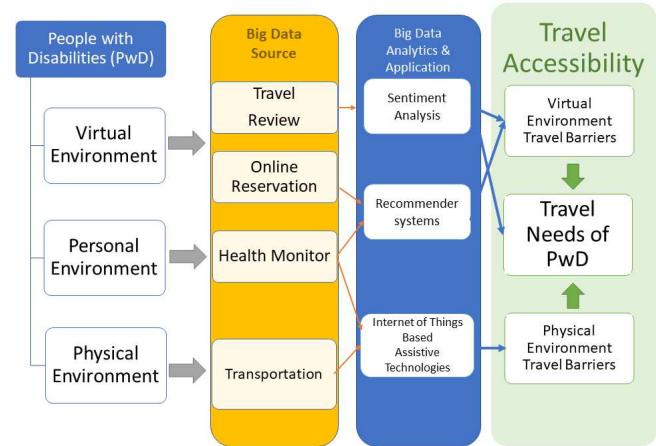
Cloud computing and big data analytics can also be helpful in health monitoring. The Artemis project mention earlier provide a example of big data analytics and cloud computing usage in health monitoring, by creating new cloud-base health analytics solutions [26]. Previous researchers have developed a mobile app to collect motion data of Parkinson's disease (PD) which is a disease resulting in mobility disorder using the smart phone 3D accelerometer and to send the data to a cloud service for storage, data processing, and PD symptoms severity estimation, which provide an user-friendly and economically affordable system to monitor and assess the condition of PD [34]. Although this system is not for people with disabilities, but it provided potentials for similar systems to be developed for different kind of disabilities.

Another application of cloud computing and big data in assistive technology is the CloudCast platform, which is a cloud-based speech recognition services that can be used for many assistive technology application for people with speech difficulties and hearing impairment, it also facilitate the collection of speech data required for the machine learning techniques [15]. Similar to Alexa Voice Service, it provide reliable speech recognition which can be used with assistive devices for people with hearing impairments, but CloudCast platform also provide customization for assistive technology applications benefiting users with speech impairment [15]. This research provided a great example of using big data and cloud computing in combine to solve a certain problem for people with disabilities (in this case it is barriers for speech impairment).

The development of information technology and assistive technology has improved the life of people with disabilities. The use disability informatics and health informatics can help researchers and service and technology providers to better understand the needs and wants for people with disabilities. Studies has discussed and proposed the great potentials to use big data source to better represent people with disability and identity and study issues and propose actions and solution to the challenges faced by people with disabilities. Using Cloud computing and big data also helps improving assistive and information technology that are now used to help people with disabilities. To improve the quality of life for people with disabilities, travel as a necessary needs and right for human cannot be ignored and the travel demands from the population with disabilities have to be addressed. Therefore, it will be beneficial and necessary to study travel accessibility with the help of big data and digital technologies, in order to improve the quality of life for people with disabilities. By reviewing previous literature, this study is exploring the potential relationship between big data and travel accessibility as shown in Figure 1.

## 4 TRAVEL ACCESSIBILITY AND BIG DATA

To explore the potential of big data applications and analytics in improving travel accessibility, the complexity of travel accessibility have to be addressed. Accessible travel includes not only the point-to-point transportation (such as air travel, flights), but also the accessibility of destination [3, 16, 30]. For people with disability to actually make the trip, they will also require booking for transportation and hotel reservation to be accessible. This study is going



**Figure 1: The Relationship Between Travel Accessibility and Big Data.**

to explore the use of big data application and analytics in different aspect of travel such as reservation, long distance transportation (in the form of airline travel), and destination transportation, the existing evidence and potentials for using these big data to improve travel accessibility..

### 4.1 Big Data and Online Reservation

Nowadays, the majority of the travel planning process happens online. Tourists would use a variety of tourism website, search engines, and reservation domains. The Internet also plays an important part at during and post travel stage, as tourists report issues on mobile applications, share traffic information in the cities, search and post on social media, take and share photos, report experience on travel websites and social media [2, 40]. These activities seems mundane and easy to complete for the general public, but for people with disabilities they can huge hassles or barriers.

Previous studies in relation to disability informatics demonstrated the profound positive impact of that the digital revolution on the life of people with disability by empowering them with the help of digital technologies [4]. However, there are still access issues in the digital world, the most urgent one is the use of the World Wide Web (WWW or Web). The Web has always had a strong awareness and been advocacy for accessibility since early on in its evolution [4]. The World Wide Web Consortium (W3C) had passed the Web Accessibility Initiative (WAI) and Web Content Accessibility Guidelines in the late 1990s [4].

A number of assistive technologies were designed to help people with disabilities to use the Web. For example BBC Education Text to Speech Internet Enhancer (BESTIE) is a CGI Perl script that can help people with disabilities who are using text-to-speech systems for Web browsing to modified the web page removing images, Java and Javascript code that may cause difficulties to understand the BBC web page content [18]. However, the limitation of BESTIE is that it is only compatible with BBC website. Other researchers also came up with Personalizable Accessible Navigation (PAN), which is a set of edge services designed to improve Web pages accessibility

which allow personalization and the opportunities to select multiple profiles, making it compatible for web as well as mobile devices [34].

The online sector of the tourism industry has quickly adopted big data applications to better understand the need of customer and to improve online experience for customers [2]. The online sector of the industry include meta-search engines (like Google), online travel agencies (like Expedia) and some information website companies that distribute tourism information (TripAdvisor)[29]. Amadeus, a tourism company known for its global distribution system, has developed a program “Amadeus Airline Cloud Availability” that can generated special result and increase search for its customers [40].

Travel domain companies like Marriott, Southwest airline, and Amtrak also developed assistive devices specially for people with disabilities to use when browsing their websites. These assistive technologies can help people with disabilities to navigate online reservation website, and help them to independently booking their travel reservations for hotels, restaurants, airplane tickets and attraction passes.

These assistive devices was design to help people with disabilities to have access to the Web. However, current web accessibility standards do not respect disability as a complex and culturally contingent interaction. The needs and demand of people with different types of disabilities are certainly not the same and this make it hard to understand and pinpoint the real needs for people with disabilities to access the Web. Researchers have proposed to use big data to better understand of the relation between disability and technology and recognized the difference of disabled people in the “Global South” where different contexts constitute different disabilities and different experiences of web access [4].

## 4.2 Accessible Transportation and Big Data

An inaccessible transport network prevents many people from going to school or studying, working, going to the doctor, meeting friends, going shopping or to the cinema and other activities that are taken for granted. However, for people with disabilities, an inaccessible transport network would left them dependent and confined in their own home [3]. More importantly an inaccessible transport network would also prevent people with disabilities to travel for business or leisure [30]. Older adults, people with disabilities, individuals in low-income households, especially those living in rural areas can face significant mobility challenges [32].

Concerns about getting into an accident, congestion, price of travel, access to transit, and lack of walkways are important issues for a large percentage of the population, but they tend to be more important for people with disabilities [32]. For today’s travel and transportation businesses, it is important to address the issue of inclusion, which is the potential to enable a broader range of people to use transportation infrastructure regardless of their individual abilities or disabilities [30]. Accessibility transportation is essential for travel accessibility, because it represent two aspect of travel accessibility: first is long distance transportation accessibility, and the second is accessible destinations. For a destination to be accessible, it have to have an accessible transport network allow people with disabilities to navigate with the destination.

**4.2.1 Airline and Big Data.** The airline industry is very familiar with big data use in their daily operations and market research. Airline companies have been using their big data which is the large volume of structured information that has been produced internally [29] to analyze prices of plane ticket. Moreover, airlines have optimized the details of planning for the crew and routing [40].

Previous studies in airline network used Big Data mined from the U.S. air transportation system over the years from 1998?2014 to characterize the network’s behavior and determine what internal and/or external drivers result in structural changes to the airline network [14]. Airline delay patents has also been studied with the help of big data by identifying by the number of late arrivals as a percent of total operations [43].

In another previous study, researchers used data from 2006 to 2008 in order to provides the result about the total flight delay for a specific period of time caused due to climate, security, carrier, National Aviation System, Arrival and Departure based on total number of flights getting delayed over in the given period of time [43]. In the study, the authors used time series analysis along with the integration of heterogeneous database to identify and achieve the Airline Seasonal Delay which is implemented and visualized in R, they were able to identify a trend line to provide the insights for the aviation industry to take future measures to avoid delays and manage them [43].

Airline studies have also used big data analytics on passenger reviews data. The advance development of social media and mobile helped the passengers to post reviews in a ubiquitous way, allow them post real time feedback over Facebook, Twitter on airports, airlines, and other travel providers [11]. However, passengers’ review can be really complex since travel activities usually involve multiple parties, therefore, the travel domain application systems are also typically managed by different stakeholders like airlines, airports, travel agencies, security and other services providers like cars, bus, trains, hotels, events. In order to provide a holistic approach to manage complex passenger reviews with data gathering, processing and disseminating, a previous study has proposed a reference architecture to manage passenger reviews where multiple stakeholders are involved by using data lakes, which can store, manage and analyze structured and unstructured data with cheaper cost, well-distributed, open sourced and powerful set of tools.

Even though previous studies on airline using big data have not address the issues of accessibility. These previous studies still show the potential for using big data source from the airline industry and passenger reviews to study the interests, motivation, needs and demands from people with disabilities.

**4.2.2 Transport Network Accessibility and Big Data.** Accessibility in the transport network studies are different defined than travel accessibility, since the urban accessibility is focus on provide access of transportation and transit to general public, not specifically people with disabilities [32]. However, since transport network accessibility usually are connected to urban transport network, which represent an important part of physical environment for people with disabilities it can provide some insight for accessible destinations especially urban destinations. Accessibility has always been a key concept in urban and regional planning for its capacity to link

the activities of people and businesses to the possibilities of reaching them effectively. The accessibility of the transport network is already challenging for general public, for people with disabilities who require special needs, it becomes a tremendous and difficult barrier, because they are required to make rapid, real-time decisions that are especially difficult for special needs populations. Therefore, previous studies using big data on urban accessibility can provide some potentials for travel accessibility studies [5].

For Urban and regional planning, accessibility represents traffic capacity to link the activities of people and business to the possibilities of reaching them effectively [32]. Accessibility is defined as "a dynamic attribute of locations that varies over time due to changes in the transport network and in the attractiveness of destinations for certain activities" [32]. Since the emergence of big data generated by social media, smart phones, satellite navigation system and other technologies, the information on transport networks has improved conclusively in recent years [5]. Navigation companies such as TomTom; websites and applications like Bing Maps, Google Map; collaborative projects like Open-Street-Map; the public availability of Transit Feed Specification and data from other transit authorities opens up a rapid growing field of research on real time and time-of-day variations in private and public transit accessibility [32]. These companies and institutions have increasingly detailed systems with information on the features of roads and public transport networks, and their databases include information on speed variations on the roads and the frequencies of passage in public transport networks, all of which contribute a more efficient and dynamic vision to urban accessibility studies [32].

Researchers have just started using these new information sources in studies on urban accessibility. Previous studies utilized data obtained from global positioning system devices to calculate speeds, congestion levels and accessibility conditions at different times of day (morning, midday, evening), other studies analyzed data from Be-Mobile system (which provided the geo-located positions of 400,000 vehicles equipped with tracking devices) to calculate car travel times [32].

Previous studies also have gathered and analyzed information from web services (such as Google map API) to calculate travel times between origins and destinations. These studies were able to use new information source and big data provided by the development of technologies to retrieve information about local locations and traffic condition for local facilities like groceries, malls, restaurants, banks, recreation centers and others, and estimate accessibility by car, walking, cycling and public transit options [32]. Previous research also used data from social media such as Twitter in combined with data from satellite navigation system (like TomTom) to provide a dynamic approach and obtain profiles that highlight the daily variations in accessibility in urban cities, identify real time influence of congestion and population location changes, by providing different accessibility profiles from different transport zone, the researchers were able to analyze the relationship between the performance of the transport network and the attractiveness of the destination [32]. Although this study is not designed to provide information for people with disability, but such dynamic approach using real time big data can also benefit people with disabilities and help them identify places to go in the city and the most accessible route to the attractions.

Another study also proposed and developed a travel assistance device for people with disabilities by using real time data from global positioning system. According to this study, recent advancements in mobile technology enabled smart phones with global positioning system provide real-time location-based services and its related data. The researchers for this study designed, implemented and tested a travel assistance device (TAD) that is designed to help transit riders with special needs using public transportation [5]. This device is a navigation software program designed to prompt individuals via a cell phone to exit the bus at a pre-set location. This travel assistance device provides the people with disabilities customized real-time audio, visual and tactile prompts for exiting the transit vehicle by announcing "Get ready" and "Pull the cord now!", based on its real-time assisted GPS data provided by the embedded GPS chip in the cell phone [5]. Once the software is downloaded and installed to the cell phone, parents, travel trainers, or other authorized individuals can access the web management page to schedule bus routes to be transmitted to the cell phone [5]. The system also provides alerts to riders, their caretakers and travel trainers when the rider with disabilities deviates from the planned route. With a website allowing easy access for the design and planning of new trip itineraries, the device allows authorized personnel (usually caregivers, family members) to monitor the rider's location in real-time from any computer [5]. The travel assistance device was catered to the needs of people with disabilities, increasing their level of independence and their care-takers security. This travel assistance device represented beneficial practice for people with disabilities [5].

However, there are still many challenges for the development of such device or services. One of the main challenges is that different needs of people with disabilities making it difficult to completely satisfy and assist people with different disabilities. Since the device proposed by this study is still at experimental stage and their test sample are limited, although through the test, the device was proved to be easy to use, it still might pose as a challenge for other people with disabilities, especially people with cognitive limitations [5].

## 5 PROMISES OF BIG DATA IN TRAVEL ACCESSIBILITY

As mentioned above, big data has some potentials in studies and research that are intended to enable and improve the ability for people with disabilities to travel independently. This section will explore different big data related technologies, applications and analytics that can help people with disabilities to travel with ease and researchers to better understand the demand and need for people with disabilities.

### 5.1 Internet of Things and Travel Assistive Technology

*5.1.1 Internet of Things Assistive Technology and Potentials for Travel use.* Assistive devices or technology (AT) for people with disabilities are not a new concept. Assistive devices refers to "any item, piece of equipment, or product system, whether acquired commercially off the shelf, modified, or customized, that is used to increase, maintain or improve functional capabilities of individuals with disabilities", for examples, canes, crutches, walkers,

wheelchairs, and shower chairs, hearings aids, visual aids, other hardware, and software that improve ICT access or communication capacities are all assistive devices [5, 41]. People with disabilities are usually seen as depend, and in need of help from caregiver or family members [49]. However, with the technological growth that has been seen in the last 20 years, a wide array of devices have been adapted, created, and utilized with the potential to create independence for individuals with disabilities. Virtual reality, sensor monitoring devices, smart phone have all been used to assist those who require assistance in their day to day lives [32].

Older adults and individuals with physical, sensory, and mental/cognitive disabilities encounter many barriers to inclusion and accessing various opportunities and services that the society has to offer [41]. In order to overcome these barriers, people with disabilities needs some forms of assistive technologies or devices. Traditional technologies has many challenges, however, with the development of technology, the Internet of Things (IoT), smart homes, smart buildings, smart cities, and other smart environments can overcome some of these challenges due to their prevalence and diverse capabilities [9].

One of human's fundamental needs is the mobility and capability of independently traveling around, including for people with physical disabilities and even blind or visually impaired individuals [5]. In order for people with visual impairment to independently travel or moving around requires indoor and outdoor navigation capabilities, the blind and visually impaired people may rely on some type of AT to supplement their navigational abilities [41]. Previous studies has proposed using the advanced sensors of the smart phones in providing meaningful interactions with the environment for individuals with different abilities [41]. A typical modern smart phone has more than twenty sensors, which include GPS-based systems that can be useful for outdoor navigation, however in general, these sensors in smart phone still may lack the required precision and reliability for use by the blind or partially sighted individuals [41].

Previous research has proposed using Bluetooth beacons and audible instructions delivered through an interface device for navigation by the blind and partially sighted people is based on the use of such as bone conduction earphones or smart phones [5, 41]. A research team has designed and tested a system "with 16 Bluetooth beacons providing pin-point accurate indoor location mapping" for unassisted mobility in the London Underground [41]. The system is based on a mobile app, Wayfnder, which is a open platform that has the promise of promoting future development on assistive devices for people with visual limitations [41]. Microsoft, Guide Dogs, UK and several other organizations have also embarked on expanding similar concepts to respond to the challenges that people with sight loss face while navigating the cities [41]. These companies and organizations has developed technology and application that allow users to start a trip and they have a "Look Ahead" and "Find the way" mode that can help people with vision limitation to explore the city and let them stop at any point and check that they are heading the right direction [5].

**5.1.2 The Big Data Challenge of IoT Based AT.** While the growth and progress the IoT and smart environments, technologies such

as sensor technologies for comprehensive monitoring and surveillance progress and advance unbelievable fast, nevertheless, there still existed many challenges for these technologies [41]. One of the most important challenges is data availability. Because these technologies generated an enormous amount of data that surpasses the processing and use capabilities [41]. For instance, real-time localization and navigation systems that are designed to assist people with visual impairment to travel around, face two major related challenges: one is the allocation of computational resources that can process the large amounts of data coming from multiple sensors and cameras, fast enough in a real-time and synchronized manner, so that they can provide real-time guidance for people with special needs [41];the second issue relates to quick and real-time access to dynamic data sets through interfaces that are appropriate for the user [41].IoT devices typically have the issues of energy constrained, with small memory, limited processing power, and restrictive communication capabilities [41]. One positive aspect of this challenge is, these dynamic data obtained from IoT based AT device are extremely valuable and could help researchers and companies to better understand the needs of people with disabilities and analyzing such data can help researchers and companies to design better product for people with disabilities [17].

Another issues that AT adoption faces is its ability of meeting the usersfi needs and desires [41]. AT has been criticized because although AT devices "fimay have technical merit, and may solve obvious problems, but still fail to address the complex interplay of issues at work and to take the most appropriate approach to dealing with these matters. Furthermore, it is important to acknowledge that there may not even be a firightfi problem to tackle. Flexibility cannot be overvalued" [41]. Due to the complexity of the needs and wants for people with different disabilities, it can be challenging to develop an assistive device that can accommodate most people, however, a holistic understanding of the intended users is required [41]. It is important for researchers and AT device engineers to understand the wants and needs of people with disabilities, to be able to design an AT product that actually can fulfill what people with disabilities want, instead of just assuming what people with disabilities needs [5].

Some people may feel intimidated by the newer technologies such as those of the IoT-based AT [15]. First these device required some sort of learning and adapting period, and for people with disabilities, it might too longer time than for "normal" people, which can be taxing for people with disabilities and give them extra pressures [41]. For people who are used to being in control of their devices, some automate processes of IoT-based AT, which was intended to provide support for people with disabilities, ironically, may pose potential stress for operating and adapting to the devices [41]. A similar issue arises from the fast pace of the development of such advanced technologies. With the rapid advancement of technology, products and service become obsolete really quickly as the newer improved version become available. As mentioned above, people with disabilities do have a learning curve and need adapting process for IoT based assistive technologies, the constant and multiple upgrades of new version can make it harder for people with disabilities to adapt. Therefore, the elderly or people with disability or dementia may miss out on obtaining the full benefits of these devices or services [5, 15, 41]. The costs, learning curves,

or simply a lack of awareness can potentially prevent these people to use new technologies at all.

With the emergence of new technologies such as Internet of Things and large scale wireless sensor system, IoT based AT emerged as potential solution and promise for improving the quality of life for people with disabilities. They provide new opportunities and can aid people with disabilities in their travel, and help them overcome travel barriers. However, there is still manage challenges for people with disabilities, especially when it come to complex situations that they are going to encounter during their travel. There is a distinct gap in IoT based AT research: the lacking of holistic understanding of the needs and wants from people with disabilities. More studies needs to be conducted on the opinions and users experience of IoT based AT.

## 5.2 Sentiment Analysis on Online Reviews

The popularity of social media, especially review sites like TripAdvisor and blogs and wikis, leads to an enormous amount of personal reviews for travel-related information on the Web [37]. More importantly, the information in these reviews is valuable to both tourists and travel and tourism practitioners for various understanding and planning processes [49]. These UGC comes from all kinds of tourists with different demographic background, within with also has reviews from people with disabilities. Therefore, analyzing hotel reviews on various website and platform that are posted by people with disabilities can help us better understand the needs of this population group. One of the most common analytics method for large amount of review data is sentiment analysis [37].

Sentiment analysis, which is also called opinion mining, is one of the most active research areas in natural language processing [37]. The aim of sentiment analysis is to define automatic tools able to extract subjective information from text in natural language, and to create structured and actionable knowledge to be used by either a decision support system or a decision maker [23, 49]. The sentiments of reviews, online reputation or online documents are usually categorized in positive, negative and (in some studies) neutral sentiments [21]. The main goal of the sentiment classification is to extract “the global sentiment based on the subjectivity and the linguistic characteristics of the words within an unstructured text” [21]. Therefore, sentiment analysis provided a framework to transform unstructured text to structured data, which make it strongly applicable to both the academic field [8]. Because of the importance of sentiment analysis to business and society, it has spread from computer science to management science and the social sciences [36]. As a social science field and business industry, tourism and travel studies have already been using sentiment analysis in the research.

Previous studies have identified two primary approaches for sentiment analysis: methods based on the combination of lexical resources and Natural Language Processing (NLP) techniques; and machine learning approaches [21]. Since 2009, researchers have been using machine learning methods in the natural language processing (support vector machine (SVM), Nave Bayes, and the N-gram model) to do sentiment analysis on TripAdvisor reviews [49]. Their study analyzed online reviews related to travel destinations, using different supervised machine learning algorithms

The algorithms to evaluate the reviews about seven popular travel destinations in Europe and North America [49].

The etBlogAnalysis project developed a combined crawler /sentiment extraction application for the tourism industry, which used a simple and robust linguistic parsing methodology with information and terminology extraction methods in order to determine relevant utterances on expression level [37]. It will also provide a warning for tourism operator such as a hotel, if too many negative entries have been generated by their reviewers [21].

In tourism studies, sentiment analysis has been compared to traditional qualitative analytic methods. A previous study compared three alternative approaches for mining consumer sentiment (manual content coding, corpus-based semantic analysis, and stance-shift analysis) from large amounts of qualitative data found in online travel reviews [9, 13]. They applied three different approaches to study consumers’ reaction to farm stays in order to demonstrate how large volumes of qualitative data can be analyzed quantitatively in a relatively efficient and reliable way [21]. Manual content coding is the same as traditional the content analysis approach involving two researchers collaborated in a manual coding process designed to extract consumer likes and dislikes from the qualitative data [20]. According to the comparison, computer generated sentiment analysis such as stance-shift analysis processing on both syntax and lexicon assures the coding maintains the statement’s context identifying what is important to the informants by the way they express their comments. Most importantly, stance-shift analysis does not categorize what the researcher thinks is important in reviewer’s words [9]. The study suggested by combining different approaches in sentiment analysis such as using stance-shift analysis first identifies the significant word segments then using corpus-based semantic analysis detects key themes in those segments helps uncover narrative themes of consumer experiences in large qualitative databases [9].

Sentiment analysis will help researchers to better understand people’s travel experience, however, there are few studies have been done to identify demographic information of the reviewer and compare the sentiment analysis result across different demographic [23]. A recent invention present the possibility of identifying demographic characteristics while conducting sentiment analysis. The invention consist of a product or service review to determine demographic information of the reviewer [6]. A sentiment text analysis is performed on the product or service review, wherein the sentiment text analysis examines the product or service review to determine a sentiment of the product or service review. The sentiment of the product or service review is categorized based on the demographic information of the reviewer [6]. This invention presents the promise of using sentiment analysis on the travel experience of people with disabilities. However, challenges still remain for research of UGC generated by people with disabilities, such as the challenge presented by privacy concerns of personal data online [27].

## 5.3 Recommender System

Nowadays tourists faces a very challenging task of trip preparation because of the huge amount of information available on the Web about tourism and leisure activities [1]. Recommender systems

becomes essential for tourists and tourism operators. For tourists, recommender systems can be a useful tools to help them make decision for travel planning, such as the choices of destinations, attractions, accommodations and restaurants. As for tourism operators, it can be a great marketing opportunities for them to reach a variety of targeted potential consumers. Complex problems such as automated planning, semantic knowledge management, group recommendation or context-awareness have by now been heavily studied in this area [31].

There are already several tourism recommender system available for general public. TIP and Heracles systems provide recommendation service through mobile devices for tourism, through implement hybrid algorithms to calculate tourist preferences, using the defined tourist profile and location data [31]. Crumpet system provides new information delivery services for a variety of different tourist population based on location aware services, personalized user interaction, accessible multimedia mobile communication that uses Multi-Agent Technology [39]. CATIS is a Web based tourist information system using context-awareness, which include context elements such as location, time of day, speed, direction of travel and personal preferences. This system provided information to tourists relevant to his or her location and time [39]. TravelWithFriends using group recommendation service, the first step is to build a recommendation list for each user and to merge them to obtain a destinations shortlist. Afterwards, each group member rates all these options and a Borda count is used to determine the best five destinations to be recommended [31].

Classical recommender systems filter the domain items according to a particular user, using his or her demographic data, past ratings or purchasing history [28]. This approach are used to recommend specific items such as books, songs or films [28]. However, it may not be suitable for travel activities, since most of time travel is an activity that involves a group of people (such as family members, friends). Therefore, it is necessary to take into account the different preferences of all members of travel group when providing recommendations [31]. Previous studies and technology reports have identified two primary options for group recommendation: the first one is to merge the lists of items recommended to each group member, or creating a group profile with everyone's preferences and then compute a single list of group recommendations [20]. The second option's first step is the same as the previous option, by constructing of a list of recommendations for each group member. In a second step though, an automatic consensus-reaching process is applied, in which individual preferences are continuously updated until a high degree of agreement between all the group members is reached [20].

The use of semantic domain knowledge in the recommendation process has heavily increased in recent years. Previous studies have defined the semantic similarity between two concepts as "the ratio between the number of different ancestors and the total number of ancestors of both concepts" [31]. The items to be recommended are clustered according to this semantic similarity and the recommendation procedure selects the best item from random clusters [39]. Previous study has shown that this procedure keeps the accuracy and increases the diversity of the results [31]. Semantic information can also be used to determine the items to be recommended in a personalized visit to a museum or destinations,

by using a shortest-path semantic distance to determine which museum objects or attractions should be recommended to the user [31].

Previous study also proposed a hybrid tourism recommendation system for persons suffering from physical or intellectual limitation. This proposed recommendation system is not simply trying to improve experience, but to create and increase the confidence of users that despite of their limitations they can visit and experience certain places without being afraid, and to help them to truly live a touristic experience. As shown in Figure 2, the system models a user stereotype profile, by identifying the user's functionality and point of interest (POI) accessibility level, which represent user's related knowledge which is layered with several knowledge representation structures and models and produce an accurate touristic recommendation plan [39]. The study represent itself as an opportunity to provide needed information to people with disabilities through a hybrid tourism recommendation system.

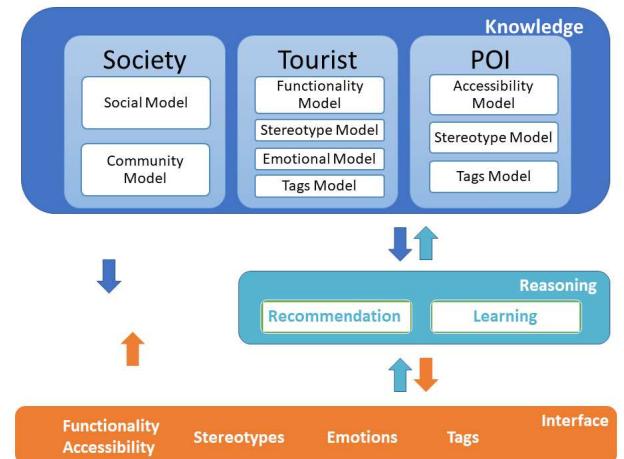


Figure 2: Hybrid Tourism Recommendation System[39].

## 6 CONCLUSION

This study has explored the big data applications and analytics in tourism industry and research, and disability related research. This study illustrated the importance of improving travel accessibility by recognizing the underestimated market for travel of people with disabilities. The lack of research on big data application and analytics in travel accessibility was identified. By recognizing the complexity of travel accessibility, this study present the potential of using big data analytics and application to better understand the need of people with disability in two travel accessibility aspects: online reservation, and accessible transportation. Although there are few studies on big data and accessible online reservations and accessible transportation directly, this study illustrate big data utilization in web accessibility, airline studies and urban accessibility. These previous studies show promises of using big data analytics and application to address accessibility issues and the needs of people with disabilities in these aspects. This study also explored the promise of big data in travel accessible by exploring:

- Potentials of Internet of Things (IoT) based assistive technology (AT): help people with disabilities overcome travel challenge presented by physical environment.
- Recommender systems: help people with disabilities to get more needed information online, and make it easier for them to navigate the virtual environment.
- Sentiment analysis on online reviews: help researchers and practitioners to better understand the needs and behaviors of people with disabilities.

However, there are still a lot challenge faced by researchers and organizations interested in improving the quality of life for people with disabilities. The most dominated challenge is the different needs for people with different disabilities types and function levels. Future studies could use sentiment analysis of reviews online generated by people with disabilities to better understand their needs and identify the differences between different disabilities groups. Future studies should also analyze dynamic data generated by the sensors on the assistive devices for people with disabilities to better understand their travel patterns and to provide more appropriate products for people with disabilities.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and TAs for i523 for his support and suggestions to write this paper.

## REFERENCES

- [1] Gediminas Adomavicius and Alexander Tuzhilin. 2015. *Context-Aware Recommender Systems*. Springer US, Boston, MA, 191–226. [https://doi.org/10.1007/978-1-4899-7637-6\\_6](https://doi.org/10.1007/978-1-4899-7637-6_6)
- [2] Rajendra Akerkar. 2012. *Big Data and Tourism*. Technical Report. Technomathematics Research Foundation.
- [3] Amadeus. 2017. *Voyage of discovery*. techreport. Amadeus, Madrid Spain. <http://www.amadeus.com> Accessed 2017.
- [4] Richard Appleyard. 2005. *Disability Informatics*. Springer New York, New York, NY, Chapter chapter 11, 129–142. <https://doi.org/10.1007/0-387-27652-1-11>
- [5] S. J. Barbeau, P. L. Winters, N. L. Georggi, and M. A. Labrador. 2010. Travel assistance device: utilising global positioning system-enabled mobile phones to aid transit riders with special needs. *IET Intelligent Transport Systems* 4, 1 (March 2010), 12–23. <https://doi.org/10.1049/iet-its.2009.0028>
- [6] D.A. Bhatt. 2014. Sentiment analysis based on demographic analysis. (May 15 2014). <https://www.google.com/patents/US20140136185> US Patent App. 13/675,653.
- [7] Ann Cameron Caldwell. 2011. *Untapped Markets in Cloud Computing: Perspectives and Profiles of Individuals with Intellectual and Developmental Disabilities and Their Families*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Chapter 30, 281–290. [https://doi.org/10.1007/978-3-642-21663-3\\_30](https://doi.org/10.1007/978-3-642-21663-3_30)
- [8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* 28, 2 (March 2013), 15–21. <https://doi.org/10.1109/MIS.2013.30>
- [9] Antonella Capriello, Peyton R. Mason, Boyd Davis, and John C. Crofts. 2013. Farm tourism experiences in travel reviews: A cross-comparison of three alternative methods for data analysis. *Journal of Business Research* 66, 6 (2013), 778 – 785. <https://doi.org/10.1016/j.jbusres.2011.09.018> International Tourism Behavior in Turbulent Times.
- [10] G. Chareyron, J. Da-Rugna, and T. Raimbault. 2014. Big data: A new challenge for tourism. In *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, DC, USA, 5–7. <https://doi.org/10.1109/BigData.2014.7004475>
- [11] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies* 68, Supplement C (2016), 285 – 299. <https://doi.org/10.1016/j.trc.2016.04.005>
- [12] Jin Chung and Dimitrios Buhalis. 2009. *Virtual travel community: bridging travellers and locals*. IGI Global, USA. 130–144 pages.
- [13] W. B. Claster, M. Cooper, and P. Sallis. 2010. Thailand – Tourism and Conflict: Modeling Sentiment from Twitter Tweets Using Naïve Bayes and Unsupervised Artificial Neural Nets. In *2010 Second International Conference on Computational Intelligence, Modelling and Simulation*. 89–94. <https://doi.org/10.1109/CIMSiM.2010.98>
- [14] E. Clemons, R. Jordan, and T. Reynolds. 2016. Airline network and competition characterization using big data approaches. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, Sacramento, CA, USA, 1–10. <https://doi.org/10.1109/DASC.2016.7777957>
- [15] Stuart Cunningham, Phil Green, Heidi Christensen, JJ Atria, A Coy, M Malavasi, L Desideri, and F Rudzicz. 2017. Cloud-Based Speech Technology for Assistive Technology Applications (CloudCAST). *Harnessing the Power of Technology to Improve Lives* 242 (2017), 322.
- [16] Simon Darcy. 2010. Inherent complexity: Disability, accessible tourism and accommodation information preferences. *Tourism Management* 31, 6 (2010), 816 – 826. <https://doi.org/10.1016/j.tourman.2009.08.010>
- [17] G Dewsbury, K Clarke, M Rouncefield, I Sommerville, B Taylor, and M Edge. 2003. Designing acceptable 'smart' home technology to support people in the home. *Technology and Disability* 15, 3 (2003), 191 – 199. <http://proxylib.uits.iu.edu/login?url=https://search-ebscohost.com.proxylib.uits.iu.edu/login.aspx?direct=true-db=ccm-AN=106746102-site=ehost-live-scope=site>
- [18] Ugo Erra, Gennaro Iaccarino, Delfina Malandrino, and Vittorio Scarano. 2007. Personalizable edge services for Web accessibility. In *Universal Access in the Information Society (W4A)*, Vol. 6. WWW2006, ACM, Edinburgh, UKfif, 285–306.
- [19] Lex Frieden. 2015. Why Disability Informatics? (02 2015). <https://sbmi.uth.edu/blog/feb-15/021115.htm>
- [20] Irma Garcia, Laura Sebastia, Eva Onaindia, and Cesar Guzman. 2009. *A Group Recommender System for Tourist Activities*. Springer Berlin Heidelberg, Berlin, Heidelberg, 26–37. [https://doi.org/10.1007/978-3-642-03964-5\\_4](https://doi.org/10.1007/978-3-642-03964-5_4)
- [21] Aitor Garca, Sean Gaines, and Maria Teresa Linaza. 2012. A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain. *E-review of Tourism Research* 10, 2 (2012), 35 – 38. <http://proxylib.uits.iu.edu/login?url=https://search-ebscohost.com.proxylib.uits.iu.edu/login.aspx?direct=true-db=hjh-AN=84339713-site=ehost-live-scope=site>
- [22] Jan Grue. 2016. The social meaning of disability: a reflection on categorisation, stigma and identity. *Sociology of Health and Illness* 38, 6 (2016), 957–964. <https://doi.org/10.1111/1467-9566.12417>
- [23] Dietmar Grbner, Markus Zanker, Gnther Fliedl, and Matthias Fuchs. 2012. Classification of Customer Reviews based on Sentiment Analysis. In *19th Conference on Information and Communication Technologies in Tourism (ENTER)*. Springer, Helsingborg, Sweden, 460–470.
- [24] Yue Guo, Stuart J. Barnes, and Qiong Jia. 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management* 59, Supplement C (2017), 467 – 483. <https://doi.org/10.1016/j.tourman.2016.09.009>
- [25] Jeffery Hoehl and Kaleb August Sieh. 2010. *Cloud Computing and Disability Communities: How Can Cloud Computing Support a More Accessible Information Age and Society?* Technical Report. Silicon Flatirons Center, Colorado, US. <https://doi.org/10.2139/ssrn.228526>
- [26] H. Khazaei, C. McGregor, M. Eklund, K. El-Khatib, and A. Thommandram. 2014. Toward a Big Data Healthcare Analytics System: A Mathematical Modeling Perspective. In *2014 IEEE World Congress on Services*. IEEE, Anchorage, AK, USA, 208–215. <https://doi.org/10.1109/SERVICES.2014.45>
- [27] Jonathan Lazar, Michael Ashley Stein, and Judy Brewer. 2017. *Disability, human rights, and information technology*. Philadelphia : University of Pennsylvania Press, [2017], Pennsylvania, USA. <http://proxylib.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true-db=cat0001ea-AN=inun.16424800-site=eds-live-scope=site>
- [28] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: A survey. *Decision Support Systems* 74, Supplement C (2015), 12 – 32. <https://doi.org/10.1016/j.dss.2015.03.008>
- [29] Shah Jahan Miah, Huy Quan Vu, John Gammack, and Michael McGrath. 2017. A Big Data Analytics Method for Tourist Behaviour Analysis. *Information and Management* 54, 6 (2017), 771 – 785. <https://doi.org/10.1016/j.im.2016.11.011> Smart Tourism: Traveler, Business, and Organizational Perspectives.
- [30] Milo N. Mladenovif. 2017. Transport justice: designing fair transportation systems. *Transport Reviews* 37, 2 (2017), 245–246. <https://doi.org/10.1080/01441647.2016.1258599>
- [31] A Moreno, L Sebastian, and P Vansteenwegen. 2015. Recommender Systems in Tourism. *IEEE Intelligent Informatics Bulletin* 16, 1 (Dec. 2015), 1–2. <http://www.comp.hkbu.edu.hk/~iib/>
- [32] Borja Moya-Gómez, María Henar Salas-Olmedo, Juan Carlos García-Palomares, and Javier Gutiérrez. 2016. Dynamic accessibility using Big Data: The role of the changing conditions of network congestion and destination attractiveness. *Networks and Spatial Economics* 1, 7 (2016), 1–18.
- [33] Open Door Organization. 2015. Open Doors Organization Market Study Press Report. (2015). <http://opendoorsnfp.org/market-studies/2015-market-study/> accessed 2017.
- [34] Di Pan, Rohit Dhall, Abraham Lieberman, and B. Diana Petitti. 2015. A Mobile Cloud-Based Parkinson's Disease Assessment System for Home-Based Monitoring. *JMIR mHealth uHealth* 3, 1 (26 Mar 2015), e29. <https://doi.org/10.2196/2196>

- mhealth.3956
- [35] Gabriel Pestre. 2016. Big Data and Disability, Part 1. Data Pop Alliance. (March 2016). <http://datapopalliance.org/big-data-and-disability-part-1/> Accessed 2017.
  - [36] Federico Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment Analysis in Social Networks*. Elsevier LTD, Oxford, Cambridge, MA.
  - [37] V. B. Raut and D. D. Londhe. 2014. Opinion Mining and Summarization of Hotel Reviews. In *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE, Bhopal, India, 556–559. <https://doi.org/10.1109/CICN.2014.126>
  - [38] Paraskovi Riga and Georgios Kouroupetroglou. 2013. Indoor Navigation and Location-Based Services for Persons with Motor Limitations. In *Disability Informatics and Web Accessibility for Motor Limitations*. IGI Global, Greece, 202–233. <https://doi.org/10.4018/978-1-4666-4442-7.ch006>
  - [39] Filipe Santos, Ana Almeida, Constantino Martins, Paulo Moura de Oliveira, and Ramiro Gonçalves. 2018. Hybrid Tourism Recommendation System Based on Functionality/Accessibility Levels. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017*, Fernando De la Prieta, Zita Vale, Luis Antunes, Tiago Pinto, Andrew T. Campbell, Vicente Julián, Antonio J.R. Neves, and María N. Moreno (Eds.). Springer International Publishing, Cham, 221–228. [https://doi.org/10.1007/978-3-319-61578-3\\_23](https://doi.org/10.1007/978-3-319-61578-3_23)
  - [40] S. Shafiei and A. R. Ghatari. 2016. Big data in tourism industry. In *2016 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC)*. IEEE, Isfahan, Iran, 1–7. <https://doi.org/10.1109/ECDC.2016.7492979>
  - [41] Seyed Shahrestani. 2017. *Internet of Things and Smart Environments*. Springer-Verlag GmbH, Cham, Switzerland.
  - [42] Ralph W. Smith. 1987. Leisure of disable tourists: Barriers to participation. *Annals of Tourism Research* 14, 3 (1987), 376 – 389. [https://doi.org/10.1016/0160-7383\(87\)90109-5](https://doi.org/10.1016/0160-7383(87)90109-5)
  - [43] M. Sornam, M. Meharunnisa, and Parthiban Nagendren. 2017. Big Data Analytics on Aviation data for the prediction of Airline Trends in Seasonal Delay. *International Journal of Advanced Research in Computer Science* 8, 5 (2017), 2248. <http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edb&AN=124636583&site=eds-live&scope=site>
  - [44] M. Viceconti, P. Hunter, and R. Hose. 2015. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics* 19, 4 (July 2015), 1209–1215. <https://doi.org/10.1109/JBHI.2015.2406883>
  - [45] N.L. Williams, A. Inversini, N. Ferdinand, and D. Buhalis. 2017. Destination eWOM: A macro and meso network approach? *Annals of Tourism Research* 64 (2017), 87–101. <https://doi.org/10.1016/j.annals.2017.02.007> cited By 0.
  - [46] Zheng Xiang, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* 44, Supplement C (2015), 120 – 130. <https://doi.org/10.1016/j.ijhm.2014.10.013>
  - [47] Karen L. Xie, Kevin Kam Fung So, and Wei Wang. 2017. Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management* 62, Supplement C (2017), 101 – 110. <https://doi.org/10.1016/j.ijhm.2016.12.004>
  - [48] WA Yasnoff. 2000. Public health informatics: improving and transforming public health in the information age. *Journal of public health management and practice* 6, 6 (11 2000), 67–75.
  - [49] Qiang Ye, Ziqiong Zhang, and Rob Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications* 36, 3, Part 2 (2009), 6527 – 6535. <https://doi.org/10.1016/j.eswa.2008.07.035>
  - [50] Ye Zhang and Shu Tian Cole. 2016. Dimensions of lodging guest satisfaction among guests with mobility challenges: A mixed-method analysis of web-based texts. *Tourism Management* 53 (2016), 13–27.

```
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 claster
Warning--empty address in claster
(There were 2 warnings)
```

```
bibtext _ label error
```

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

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

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

```
[2017-12-10 13.49.44] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
```

```
Missing character: ""
Typesetting of "report.tex" completed in 1.3s.
```

---

## Compliance Report

---

```
name: Weixuan Wang
hid: 234
paper1: Oct 22 2017 100%
paper2: Nov 9 2017 100%
project: 100% Dec 07 17
```

```
yamlcheck
```

---

```
wordcount
```

---

```
11
wc 234 project 11 9039 content.tex
wc 234 project 11 10088 report.pdf
wc 234 project 11 4940 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  
-----

320: the potential relationship between big data and travel accessibility  
as shown in Figure \ref{F:present}.

322: \begin{figure}[htb]

323: \centering\includegraphics[width=\columnwidth]{images/present.png}  
}

324: \caption{The Relationship Between Travel Accessibility and Big  
Data.}\label{F:present}

846: As shown in Figure \ref{F:rec}, the system models a user  
stereotype profile, by

853: \begin{figure}[htb]

854: \centering\includegraphics[width=\columnwidth]{images/rec.png}

855: \caption{Hybrid Tourism Recommendation  
System}\cite{Santos2018}\label{F:rec}

figures 2

tables 0

includegraphics 2

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 >= includographics)
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--empty publisher in cluster
Warning--empty address in cluster
(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 8211  
non ascii found 8217  
non ascii found 8217  
non ascii found 8230  
non ascii found 8216  
non ascii found 8217  
non ascii found 239  
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 in predict house price

Yujie Wu

Indiana University Bloomington

Bloomington, Indiana 47401

yujiwu@iu.edu

## ABSTRACT

House price changes dynamically and it costs human power to evaluate the price. This project uses a large data set obtained from house dealings and algorithms to predict if house price is reasonable.

## KEYWORDS

i523, HID235, House Price, Logistic Regression, linear Regression

## 1 INTRODUCTION

Asking a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this project is aimed to overcome the challenges of predicting the final price of each home.

## 2 LINEAR REGRESSION

In statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable  $y$  and one or more explanatory variables (or independent variables) denoted  $X$ . The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression[3].

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of  $y$  given the value of  $X$  is assumed to be an affine function of  $X$ ; less commonly, the median or some other quantile of the conditional distribution of  $y$  given  $X$  is expressed as a linear function of  $X$ . Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of  $y$  given  $X$ , rather than on the joint probability distribution of  $y$  and  $X$ , which is the domain of multivariate analysis[3].

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than

models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine[3].

## 3 LOGISTIC REGRESSION

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analyzed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model[2].

Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor[2].

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict whether a patient has a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient. Another example might be to predict whether an American voter will vote Democratic or Republican, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic

regression to sequential data, are used in natural language processing[2].

## 4 EXPERIMENT

```
In [1]: import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.metrics import accuracy_score
In [2]: from sklearn.ensemble import RandomForestClassifier
import pandas as pd
from sklearn.preprocessing import LabelEncoder
import numpy as np
df=pd.read_csv('C:\\\\Users\\\\Yujie\\\\s
VMware\\\\Desktop\\\\training_data.csv')
df.head()
```

	Id	MSSubClass	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2
0	1	60	8450	7	5	2003	2003	196	706	0
1	2	20	9600	6	8	1976	1976	0	978	0
2	3	60	11250	7	5	2001	2002	162	486	0
3	4	70	9550	7	5	1915	1970	0	216	0
4	5	60	14260	8	5	2000	2000	350	655	0

	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	MiscVal	MoSold	YrsSold	SalePrice
0	61	0	0	0	0	0	0	2	2008	208500
298	0	0	0	0	0	0	0	5	2007	181500
0	42	0	0	0	0	0	0	9	2008	223500
0	35	272	0	0	0	0	0	2	2006	140000
192	84	0	0	0	0	0	0	12	2008	250000

```
# clean data and show del df['error']
Out[26]:
df.head()
```

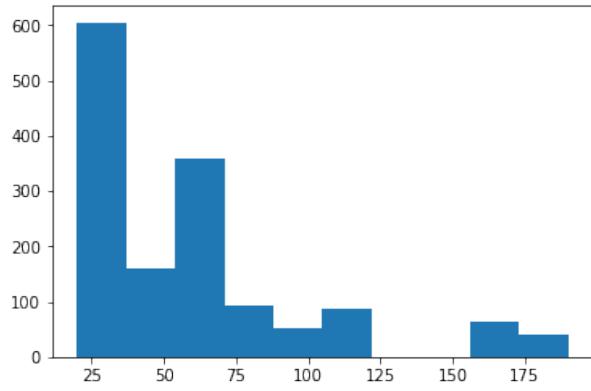
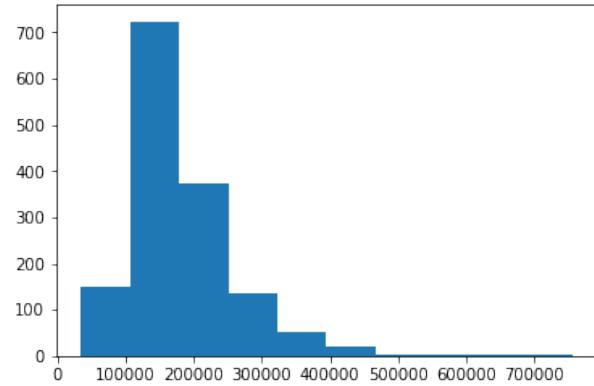
```
# clean data and show del df['error']
Out[26]:
df.head()
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig
0	50	RM	L70	L10517	Pave	None	Reg	Lvl	AllPub	Inside
1	20	RL	L70	L10517	Pave	None	IR1	Lvl	AllPub	Inside
2	20	RL	L70	L10517	Pave	None	Reg	Lvl	AllPub	Corner
3	20	RL	G70	L10517	Pave	None	Reg	Lvl	AllPub	Inside
4	20	RL	G70	L10517	Pave	None	IR1	Lvl	AllPub	Inside

	GarageCars	GarageQual	GarageCond	PavedDrive	PoolArea	PoolQC	Fence	SaleType	SaleCondition	binSalePrice
2	Fa	TA	Y	N	None	None	WD	Abnorml	N	
1	TA	TA	Y	N	None	MnPrv	WD	Normal	N	
2	TA	TA	Y	N	None	None	WD	Normal	N	
2	TA	TA	Y	N	None	GdWo	WD	Abnorml	N	
2	TA	TA	Y	N	None	None	WD	Normal	Y	

```
In [50]: # show the distribution of a column col =
list(df['SalePrice'])
plt.hist(col)
plt.show()
# show the distribution of another column
col = list(df['MSSubClass'])
plt.hist(col)
plt.show()
```



### 0.0.1 Splitting into train and test

```
In [6]: X =
df[['MSSubClass','LotArea','OverallQual','OverallCond','YearBuilt',
'YearRemodAdd','
y = df['SalePrice']

3
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=3)

In [7]: print("X_train: ")
print(X_train.shape)
print("y_train: ")
print(y_train.shape)
print("X_test: ")
print(X_test.shape)
print("y_test: ")
print(y_test.shape)
```

X\_train:

(1168, 18)

y\_train:

(1168,)

X\_test:

(292, 18)

y\_test:

(292,)

```

0.0.2 Using Linear Regression
In [53]: # import model
from sklearn.linear_model import LinearRegression
# instantiate
linreg = LinearRegression()
# fit the model to the training data (learn the
# coefficients)
linreg.fit(X_train, y_train)
Out[53]: LinearRegression(copy_X=True, fit_intercept=True,
n_jobs=1, normalize=False)

```

---

```

0.0.3 show the coefficients of the linear regression model
In [54]: # print the intercept and coefficients
print(linreg.intercept_)
print(linreg.coef_)
1624810.78299
[-1.32598253e+02
 7.42385752e+01
 4.81195345e+00
1.17264859e+00
3.87166477e+02
3.94460800e+01
3.27974313e+04
6.87797648e+01
8.24586253e+01
3.52017131e+02
3.05508150e+01
2.55083195e+01
3.67581004e+01
6.78306474e+01
1.90173648e+02
-2.30935682e+00
-7.58643734e+02
-1.29045399e+03]

```

---

```

In [57]: # pair the feature names with the coefficients
list(zip(X, linreg.coef_))
Out[57]: [('MSSubClass', -132.59825288297418),
('LotArea', 1.1726485927894912),
('OverallQual', 32797.431340275551),
('OverallCond', 352.01713092259615),
('YearBuilt', 74.238575183515508),
('YearRemodAdd', 387.16647668821042),
('MasVnrArea', 68.779764836201025),
('BsmtFinSF1', 30.550814975040396),
('BsmtFinSF2', 4.811953445828749),
('WoodDeckSF', 39.446080031209931),
('OpenPorchSF', 82.458625258585016),
('EnclosedPorch', 25.508319545409904),
('3SsnPorch', 36.758100373945126),
('ScreenPorch', 67.830647415049953),
('PoolArea', 190.17364768682),
('MiscVal', -2.309356818373999),
('MoSold', -758.64373443570901),
('YrSold', -1290.4539903809468)]

```

---

```

0.0.4 Make Prediction
In [58]: y_pred = linreg.predict(X_test)
print(y_pred)

```

```

[ 93201.65287675 142554.76960023 207127.40186274
 129212.10104713
284603.13460876 65916.12615103 200722.35383783
169593.14341713
138022.19789723 138387.20437863 143561.19507458
150824.4043784
141397.06769187 242451.92943507 294242.12637714
155754.66681213
137312.71913462 129724.80248527 114876.63835247
150591.3640421
123366.39990094 150807.12337525 243530.22799299
80763.41783606
197680.6442071 103336.1791911 164351.0638006
90905.30200621
168564.67988019 115830.7939216 155494.23908365
100141.70869163
129753.99108892 156909.62744952 198024.33401226
127554.26516619
202396.44876193 110389.71894826 176051.07333227
145661.99691863
220408.37444225 116538.72044784 244706.34275193
195197.31242857
278775.03276079 152402.88891841 144044.1694957
99239.71058997
131931.4514502 110682.0707577 286158.2247229
135834.2312254
97216.99608222 143430.65499509 119919.1222882
142851.58478543
204412.40936722 153442.22728046 132396.82935032
257628.30990488
169933.91094914 120347.46703607 105703.73721332
334713.28117372
114893.96280934 234743.60884166 183600.27577235
258490.04570492
375690.10951031 111056.04841375 172340.59505253
136413.8949467
144985.53609179 223770.11250678 153958.15058402
118182.70605492
231712.95504903 102900.49898239 154389.2668046
127207.38840213
145004.55720471 176843.72114247 169051.0221719
148482.89114126
5
313474.32652747 152493.90863482 173352.98469532
126431.410949
706997.59838104 273836.15113809 408289.15199464
184268.88377474
176636.23715661 275235.88756875 209025.79856597
301622.54866502
153471.41164935 210322.3456924 140373.14967059
147240.88073756
179983.14477495 236775.79874229 144459.66092891
253623.49760802
176613.8745506 229147.89622806 185611.33497853
184942.64455848
68523.77301979 111428.57621911 95690.85464389
210924.92842256
236920.78986689 179625.99265486 146998.08010809
218710.14496723
128961.12389426 144822.81021683 180904.35339614
126176.27283423

```

```

99200.6426002 257960.86042887 200015.00187306
189350.38765334
161309.58475033 188509.81215872 207842.304773
133745.60572444
146175.12018952 311506.70403465 123810.46022005
178891.00062911
161368.3460505 141225.57119693 247990.6084084
134197.93148368
105114.81700542 66238.30229136 178481.22924629
169327.77885027
115722.77661632 163980.78329632 259041.24033003
263441.73011914
283269.97449944 291907.40480834 197028.67375584
253486.2658659
104257.64311109 332402.81936497 235123.35447297
117530.28281743
331867.34936129 166007.48240655 203642.02685762
366937.27511738
254191.87188157 124456.0621068 91276.03127415
115556.73782587
135405.78846278 218110.30607514 214268.7945742
140032.81401882
182040.72686091 176083.82849172 189209.9335995
116292.72416312
156432.52235127 204191.85524977 243977.32691289
217247.92836948
285376.77266393 246843.94978091 222100.61618257
167537.81203899
173903.04681628 161022.49525822 142938.09421832
236955.86271045
118908.1141022 191184.99717802 264889.79476426
125154.14702448
227251.47377346 237911.61180972 161841.30235331
191446.14163809
116625.79736531 145735.12537565 105372.79335894
316582.43267848
122028.63732317 290593.52200424 185714.04747969
190352.49050499
231514.01368592 230079.79348368 127676.13717555
211300.51925447
195483.82714307 192652.60353829 165237.26624796
120930.29413027
246160.5957099 281456.29298689 246797.10864379
119311.41048692
125885.59543623 237312.78172775 80648.88617473
121685.26043487
114935.23554672 161184.27138098 148054.27442727
131650.06362918
201165.65733849 296531.91858395 130462.5135931
249027.70239568
177623.04403208 166277.96320103 94660.31809216
72766.55532673
120114.744019 102331.03622172 132734.52163217
195045.15386957
206238.5262443 254559.77871816 103180.37264793
217595.03298126
200740.09709761 309978.95599912 120679.0962053
105614.64472826
246644.91857921 132884.29067159 333136.91588735
167381.82062275
160037.46133837 280268.75854488 166764.14578859
163691.73116515

```

```

172120.87702276 86744.64311271 159726.64897898
152045.32160112
148376.7717957 261196.38354104 190853.69410498
147271.05605036
126117.61348537 122639.71249082 227607.46595295
170013.76929953
136917.40993332 157131.07140997 127090.52733891
135446.00011116
271380.56027395 138338.59566889 233995.26083728
141733.47586868
138265.15523182 200673.53942787 186815.23291077
226030.65231941
197560.45885543 392086.99980809 199746.97825355
194749.30417935
122346.88308178 170839.72571199 37264.46454778
153581.22281016
6
183230.11599231 143418.12876415 170655.31440413
60936.55821163
279635.84539619 141986.8959917 163848.54119987
130962.64226048
194210.12078339 222988.18563234 252212.98444331
371404.24966183
219031.5941132 63288.51714085 112594.85348706
139413.4862122 ]

```

0.5 Finally Check Error 0.6 Here, we use Root Mean Squared Error (RMSE) which is the square root of the mean of the squared errors for error checking:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

```
In [59]: from sklearn import metrics
        import numpy as np
        print(np.sqrt(metrics.mean_squared_error(y_test,
                                                y_pred)))
46554.0967726
```

```
0.0.7 Using logistic Regression
In [8]: from sklearn.linear_model import LogisticRegression
        logreg = LogisticRegression()
        logreg.fit(X_train, y_train)
        y_pred = logreg.predict(X_test)
        print(y_pred)
```

```
[135000 139400 215000 85000 259500 108480 161750 93000
103600 129500
80000 112000 115000 194000 171000 145000 130250 130500
151000 395192
135000 266500 335000 135000 197000 110000 116000 135000
139000 98000
181000 52500 110000 161000 248328 149000 215000 112500
143000 138000
224900 86000 215000 162900 190000 165500 79000 87000
135000 112000
223000 85400 90000 145000 143000 129500 204900 145000
127500 275000]
```

```

194000 114500 157900 260400 143500 254900 180000 190000
223000 150000
215000 89500 145000 276000 111250 220000 172500 118964
130500 127000
137000 146000 196000 123600 380000 266500 229000 129900
324000 181000
260400 144152 202500 279500 319000 190000 130500 167000
171000 140000
142000 250000 98000 275000 79900 215000 193000 199900
139600 143000
137450 185900 328900 215000 259500 205950 87000 143900
167240 127000
101800 224900 196500 167000 108959 180000 162000 129900
176000 319000
124500 155900 266000 106000 215000 147000 121600 110000
165000 176000
79900 181000 195400 290000 315750 260400 153900 325624
112500 293077
181000 141000 342643 139000 135000 325000 174000 82500
78000 131500
120500 446261 378500 135000 191000 177500 79000 80000
258000 155900
263000 220000 378500 165000 184900 214000 128500 165500
159000 185900
160000 146000 240000 131500 328900 239000 166000 191000
80000 118000
7
81000 325000 117500 222000 136905 202665 220000 266000
109500 244600
181000 159500 140000 98000 195400 184900 378500 110000
200000 278000
161500 135000 180500 144152 190000 114500 181134 297000
147000 143750
155900 142000 107500 99500 117500 116050 115000 181000
215000 203000
122000 138800 104900 315000 136905 169000 250000 149000
305900 149000
164900 229000 181000 158000 144000 125500 145000 243000
130500 250000
179900 158000 145000 117500 187500 145000 117000 130000
135000 99500
215000 89000 143000 189950 154000 145000 163000 200000
181000 264132
150000 289000 167000 137900 67000 145000 127500 141500
214000 151000
209500 118000 161500 127500 143250 230000 290000 485000
191000 85000
110000 173000]

```

---

Classification accuracy:

Proportion of correct predictions

Common evaluation metric `for` classification problems

In [10]: `from sklearn import metrics`

`import numpy as np`

`print(metrics.accuracy_score(y_test, y_pred))`

0.0

---

## 5 ALGORITHM COMPARISON AND ANALYTICS

Linear Regression is used to establish a relationship between Dependent and Independent variables, which is useful in estimating the resultant dependent variable in case independent variable change. For example, using a Linear Regression, the relationship between Rain (R) and Umbrella Sales (U) is found to be -  $U = 2R + 5000$ [1]. This equation says that for every 1mm of Rain, there is a demand for 5002 umbrellas. So, using Simple Regression, you can estimate the value of your variable.

Logistic Regression on the other hand is used to ascertain the probability of an event. And this event is captured in binary format, i.e. 0 or 1. For example, I want to ascertain if a customer will buy my product or not. For this, I would run a Logistic Regression on the (relevant) data and my dependent variable would be a binary variable (1=Yes; 0=No).

In terms of graphical representation, Linear Regression gives a linear line as an output, once the values are plotted on the graph. Whereas, the logistic regression gives an S-shaped line

The logistic model is unavoidable if it fits the data much better than the linear model. And sometimes it does. But in many situations the linear model fits just as well, or almost as well, as the logistic model. In fact, in many situations, the linear and logistic model give results that are practically indistinguishable except that the logistic estimates are harder to interpret[1].

For the logistic model to fit better than the linear model, it must be the case that the log odds are a linear function of X, but the probability is not. And for that to be true, the relationship between the probability and the log odds must itself be nonlinear. But how nonlinear is the relationship between probability and log odds? If the probability is between 0.20 and 0.80, then the log odds are almost a linear function of the probability[1].

Interpretability is not the only advantage of the linear probability model. Another advantage is computing speed. Fitting a logistic model is inherently slower because the model is fit by an iterative process of maximum likelihood[1]. The slowness of logistic regression isn't noticeable if you are fitting a simple model to a small or moderate-sized dataset. But if you are fitting a very complicated model or a very large data set, logistic regression can be frustratingly slow.

The linear probability model is fast by comparison because it can be estimated noniteratively using ordinary least squares (OLS). OLS ignores the fact that the linear probability model is heteroskedastic with residual variance  $p(1-p)$ , but the heteroscedasticity is minor if p is between 0.20 and 0.80, which is the situation where I recommend using the linear probability model at all. OLS estimates can be improved by using heteroscedasticity-consistent standard

errors or weighted least squares[1]. In my experience these improvements make little difference, but they are quick and reassuring.

## REFERENCES

- [1] 2009. VMware vSphere, the First Cloud Operating System, Provides an Evolutionary, Non-disruptive Path to Cloud Computing white paper. Online. (2009). <http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/cloud/vmw-09q2-white-paper-cloud-os-p8-r1.pdf>
- [2] Alessandro Duminuco Nico Janssens Thanos Stathopoulos Fabio Pianese, Peter Bosch and Moritz Steiner. 2010. Toward a Cloud Operating System. Online. (2010). <http://ieeexplore.ieee.org/abstract/document/5486552/>
- [3] Lama A. Aladro Hodan M. Musse. 2016. Cloud Computing: Architecture and Operating System. Online. (7 2016). <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7976640>

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  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst  
Warning--no key, author in C3  
Warning--no author, editor, organization, or key in C3  
Warning--to sort, need author or key in C3  
Warning--no key, author in C3  
Warning--no key, author in C3  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst  
Warning--no key, author in C3  
Warning--no author, editor, organization, or key in C3  
Warning--empty author in C3  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst

```
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst  
(There were 16 error messages)  
make[2]: *** [bibtex] Error 2
```

```
latex report
```

```
=====
```

```
[2017-12-10 13.49.51] 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 1.4s.  
./README.yml  
 9:81     error    line too long (86 > 80 characters)  (line-length)  
 19:1     error    trailing spaces  (trailing-spaces)  
 23:81    error    line too long (89 > 80 characters)  (line-length)  
 23:89    error    trailing spaces  (trailing-spaces)  
 24:77    error    trailing spaces  (trailing-spaces)  
 25:81    error    line too long (106 > 80 characters) (line-length)  
 25:106   error    trailing spaces  (trailing-spaces)  
 26:81    error    line too long (109 > 80 characters) (line-length)  
 26:109   error    trailing spaces  (trailing-spaces)  
 38:81    error    line too long (88 > 80 characters)  (line-length)  
 38:88    error    trailing spaces  (trailing-spaces)  
 39:81    error    line too long (87 > 80 characters)  (line-length)  
 39:87    error    trailing spaces  (trailing-spaces)  
 40:49    error    trailing spaces  (trailing-spaces)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Wu, Yujie  
hid: 235  
paper1: 100%, 10/27/2017  
paper2: 100%, 11/06/2017
```

```
yamlcheck
```

```
-----
```

```
wordcount
```

6

```
wc 235 project 6 2296 report.tex  
wc 235 project 6 2236 report.pdf  
wc 235 project 6 95 report.bib
```

find "

---

84: \par In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analyzed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model\cite{C2}.

142: In [7]: print("X\_train: ")

144: print("y\_train: ")

146: print("X\_test: ")

148: print("y\_test: ")

passed: False

find footnote

---

42: \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
```

---

```
104: \includegraphics[width=0.95\columnwidth]{3}
106: \includegraphics[width=0.95\columnwidth]{4}
117: \includegraphics[width=0.95\columnwidth]{1}
119: \includegraphics[width=0.95\columnwidth]{2}
131: \includegraphics[width=0.95\columnwidth]{output_3_0.png}
133: \includegraphics[width=0.95\columnwidth]{output_3_1.png}
```

```
figures 0
```

```
tables 0
```

```
includegraphics 6
```

```
labels 0
```

```
refs 0
```

```
floats 0
```

```
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 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

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3085 of file ACM-Reference-Format.bst

Warning--no key, author in C3

Warning--no author, editor, organization, or key in C3

Warning--to sort, need author or key in C3

Warning--no key, author in C3

Warning--no key, author in C3

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3131 of file ACM-Reference-Format.bst

Warning--no key, author in C3

Warning--no author, editor, organization, or key in C3

Warning--empty author in C3

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans  
while executing---line 3229 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans

```
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Fabio Pianese, Peter Bosch, Alessandro Duminuco, Nico Jans
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 16 error messages)
```

```
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
```

# Importance of Big data in predicting stock returns and price

Gagan Arora  
Indiana University  
2709 E 10th St  
Bloomington, Indiana 47401  
gkarora@iu.edu

## ABSTRACT

In this project, we will discuss the importance of big data in finance industry in predicting financial stock values. We will be using python libraries to fetch financial data from yahoo finance and will further predict the stock price returns of few selected technology companies such as Amazon, Yahoo depending on the historical data of x[16] years. Similarly, we will predict the returns based on y[10] years of data. The prediction will be based on SP 500 market return and market risk volatility. Here y is greater than x and then we will compare the predicted returns with the current returns. For the comparison we will be using the testing time frame as mentioned in the project later. This project will help us understand if more historic data helps in predicting the stock price returns or it adds noise. We will be using statistical approach and CAPM [capital asset pricing model ] to predict stock price. Analysis will be done on the jupyter notebook

## KEYWORDS

HID-301, Stock Price prediction, stock returns, SP500, risk free market, CAPM model, root mean square analysis, stock beta, Finance, Statistics,mean, variance, market premium, python,yahoo finance, i523

## 1 INTRODUCTION

By its nature of the business, the finance industry is always driven and dominated by data. The existence of Big data in the finance industry has exposed the big opportunity of growth and value extraction but at the same time imposed the various new challenges, which demand new skill set. [5] suggests that finance experts believe there is a huge potential in terms of value extraction from the financial big data. They also believe that finance industry can benefit more than any other industry. Historically, data was always there in some format either non-digital or digital. However, with digitalization, this data has fallen into the prevalence of high volume of information, which we call as Big Data. Dominant drivers for the actuality of big data in the finance industry are mainly customer call logs, social media, news feed, regulatory data etc. Call logs, news feed and etc. fall into the category of unstructured data which is identified as an area where we can extract vast amount of business value.

[4] talks about the three V of big data in finance industry: volume, velocity and variety. [6] clearly depicts the amount of financial data pouring in the daily basis. TechNaviofis forecast (Technavio 2016) predicts data will grow at a CAGR [compound annual growth rate] of 61 percent over the period of 2017-2021. According to the IDC financial insight 2016, every second there is around 10,000-payment

card transaction and this number is expected to double by the end of this decade. The Capgemini/RBS Global payments study for 2012 suggests there was about 260 billion transactions in 2012 and is expected to grow between 15 and 22 percent for developing countries. Main drivers contributing to the big data in the finance industry are Data growth, increasing scrutiny from regulators, digitalization of financial products, changing the business model and increased customer insight platforms such as customer service. [4] shows 76 percent of banks say the business driver for embracing big data is to enhance customer engagement, retention, and loyalty and seventy one percent of banks say that to increase their revenue, they need to better understand customers and big data will help them to do so.

Thinking about the data strategy, the financial industry has taken the business-driven approach to a big data. According to the IBM report, all financial organizations are not keeping the same pace as peer industry is keeping. Today because of increased competition, customers always expect more personalized banking service and at the same time, there is increased regulatory surveillance which in result creates big pressure on finance industry to better utilize the value of Big data. To achieve better-personalized experience, many banks have started the initiative to utilize the information gained from the vast ocean of data to offer better-personalized products and gain competitive advantage. Despite the fact that financial industry is data-driven, there is a gap in the amount of initiative financial industry has taken to extract the value out of big financial data. Technavio 2016 report has shown only 26 percent of financial organizations has focused on understanding the principal notation of Big data and most of those 26 percent are still struggling to define the clear roadmap. This clearly concludes that finance industry lag behind their cross-industry peers in using more varied data types. A good example to support this fact is that there are very less research and domain knowledge in extracting value out of retail bank call logs.

Big data technologies not only help in extracting the effective business value but analysis of unstructured data in conjunction with a wide variety of data set also helps in extracting commercial value. Big data in finance industry does not necessarily decode to valuable or actionable information. The real benefit lies in developing the technologies, which can be used to extract business and commercial value. [15] talks about what all advantage we can extract from the big data in the finance industry. Few examples are: Detection of false rumors that try to manipulate the finance market, Assessment of exposure to a reputational risk connected to consulting service offered by banks to their customer and Discover topic trends, detect events, or support the portfolio optimization

or asset allocation. Big data based pattern recognition can also help in enhanced fraud detection systems and prevention capability systems. Other benefits of utilizing big data include building a machine learning based algorithm to achieve higher performance and accuracy in the trading algorithm and Enhanced market trading analysis. There has been proven research [12] which states more data increases accuracy and precision of simulations which is the backbone of financial modeling based analytics. This research [12] states modern modeling techniques are data hungry. In this project we will extract inference if more financial data can be used to have better prediction.

## 2 USE OF STRUCTURED FINANCIAL DATA

This reflects the data which has a higher degree of an organization such as a relational database where information/data is easily searchable and we can easily apply standard algorithm to extract patterns out of it. In this project we will be using Yahoo finance structured data. Examples of such data set include yahoo financial data, trading applications, enterprise finance resource planner, Retail banking systems, Credit history database systems and other financial applications that use legacy application systems. Structured data always has a big advantage of being easily entered, stored, queried and analyzed. Most of the personal banking financial statements are stored in a structured way. Structured dataset combined with the distributed systems can be leveraged to achieve structured big data set on which we can run optimized SQL queries to retrieve patterns. [9] discusses various SQL based ways to specify information quality in data which can be used to filter out the noise. In this project we will be using structured data.

## 3 VARIOUS CHALLENGES UTILIZING BIG DATA VALUE IN FINANCE INDUSTRY

There are multiple challenges and constraints in extracting value out of big financial data. The biggest challenge is old IT culture and infrastructure. The much financial organization still uses old IT infrastructure which is not compatible with the big data application thus fail to take advantage of big data. Other challenges include lack of skill set and data privacy and security. With the emergence of digitalization, customer data is saved persistently because of which there has been continued concern regarding the customer privacy. Regulatory bodies guidelines on customer data are always ill-defined because of which there is always a concern regarding the use of customer data. In this project we will use standard python libraries to fetch financial data from yahoo finance. Analysis will be done on the jupyter notebook.

## 4 STOCK RETURNS PREDICTION - LITERATURE REVIEW

Authors of [1] discuss the importance of stock price and returns prediction based on the data extraction of historic data. This research [1] also shows historic financial data has definitive predictive relationship to the future value of stocks. Stock prediction always help investors to decide perfect timing of buying or selling stocks. There are various data mining, artificial neural networks and machine learning techniques available for the stock price prediction based on the value extraction from the historic financial data. Based

on the complexity of stock price matrix, pricing mechanism is essentially a non linear complex system. Authors of [14] and [13] state many predictive algorithm is based on the fundamental analysis of macroeconomics and company fundamentals. [11] states problem with the fundamental analysis is that it is too much focused on the intrinsic and lacks the quantitative aspect of the historic financial data. On the broad category we can define stock prediction analysis is based on two types of analysis: qualitative and quantitative. Choice of analysis is mainly based on the fact if we want to have short term analysis or long term analysis. In this project we have ten and sixteen years of training data and used close to one year of testing data. Since our analysis is based on the historic data we have chosen to do quantitative analysis. Quantitative analysis is based on the pattern extraction, fact that history repeats and future financial drivers can be extracted based on the historic data. Advantage of using quantitative analysis is that we can use statistical confidence interval to validate the analysis.

There is a huge benefit of using machine learning algorithms in predicting stock prices. These algorithms made easy to cope up with the various financial events such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc. In this project we have used hybrid approach of combined CAPM [Capital asset pricing] model and machine learning algorithm to mine data of sixteen and ten years of data and used close to 1 year of testing data. These machine learning algorithms can further be used to predict various financial events. Other approaches suchs as neural networks algorithms, SVM, logistic regression and multiple descriminator analysis can also be used to predict financial events. Example, [2] in their research they proved neural networks algorithms performs better in predicting financial events as compared to multiple descriminator analysis. There are other applications which use these algorithm to find predicted credit rating of a company. Credit rating plays a very important role doing qualitative analysis of the financial health of a company. On the other hand, accuracy of these algorithm is a big challenge because of the amount of huge data which it uses as input. Typically, accuracy of these algorithms is validated based on square root method.

In this project we have used several years of data for analysis which involves more than hundred thousands of rows with multiple columns. Then this data is analyzed two dimensionally with the same set of market return rows. Since this analysis is calculation intense, In the end we also have performed root mean square analysis.

Over the past few years there has been drastic changes in the way stock market operates. With the emergence of advance web services, there has been powerful enhancement in the data communication between various financial application. Because of which there is ocean of real time data is available, thus machine learning algorithm, neural networks algorithms, SVM, logistic regression and multiple descriminator analysis needs to be smarter. Forecasting stocks and financial parameter is of great interest to the investors. As discussed earlier these algorithms needs to modified depending on the fact if we want to have short term profit or long term profit.

[8] has shown the very interesting analysis of comparing the prediction of stock market with the random walk hypothesis. Author of [8] ran an experiment in which he tossed a coin and recordes

the results and mapped head with the company profit and tail with the company loss. Then result of this experiment was shown to the investors pretending these are the actual market profit and loss. Looking at the result graphs, investors believed it as a actual prediction. This research has shown the altogether different outlook which states stock price prediction and forecast can be fooled and stock prices are perfectly random in nature. On this theory many researchers have classified profit based on three hypothesis:

- Weak form Efficient Market Hypothesis: The weak form of the hypothesis states one can not generate profit by just looking at patterns and trends of stock market.
- Semi Strong Efficient Market Hypothesis: The semi strong form of the hypothesis states only possible way of generating profit is via inside trading.
- Strong form Efficient Market Hypothesis: The strong strong form of the hypothesis states its not possible to generate profit since stock market behaves in perfect random way.

However, if we are running root mean square analysis we can surely compare the accuracy of various algorithm and arrive at conclusion which algorithm is viable for prediction.

## 5 FINANCIAL DATA EXTRACTION

In this section, we will discuss various technical requirements needed to achieve value extraction from the big data in the finance industry. There are various technical requirements such as data Acquisition, data quality, data extraction, data integration, decision support. In order to fulfill requirements, a hybrid approach combining computer science, algorithms, statistics, data mining, machine learning and pattern recognition study needs to be adopted. To explore the advantage of big data there have been initiatives like data virtualization, multi-document summarization, pattern recognition from LOGS and many start-ups have been emerged. All big companies such as Microsoft, Google, IBM and Amazon are investing heavily in this field to leverage business and commercial value out of it. There has been changed in the industry pattern where financial industry is resorting big data to strategize their business. According to [6] with a very rapid pace, the financial industry is utilizing big data advantage in investment analysis, econometrics, risk assessment, fraud detection, trading, customer interaction analysis and behavior modeling. If we look at the Big promise the Big data holds in the finance industry, progress in this field is still in nascent stage and we expect more growth in upcoming years. In this project we will discuss jupyter notebook based solution for Data extraction.

In this project we have used jupyter notebook and rich python libraries to fetch financial stock data. Later in this paper we will discuss the stock data extraction in detail. Later we will also discuss what are different ways to fetch stock data and will discuss few important functions which python libraries

## 6 FETCHING FINANCIAL STOCK DATA

Fetching structured precise data is always a challenge. There are different ways to fetch the stock market data. In this project we will be fetching data from yahoo finance via python libraries which internally makes remote web service call to the yahoo webserver. There are also other ways to fetch data such as:

- Direct download of csv files from yahoo finance or google websites.
- Make web api call to download the data in the json/XML format
- Use python libraries to download data, which internally makes remote web service call to the yahoo webserver. This is preferred way of doing since it allows you to save data to system variables directly.
- Call yahoo or finance web service from the application.
- Calling VBA function in excel to fetch yahoo stock data
- Quandl best for using core financial data and this website also includes access to rich python libraries.
- Google sheet has feature to fetch real time stock prices
- Install stocks macros in excel

In this project we have exhaustively used python for data manipulation. Reasons for using python are:

- Syntax is super easy which comes with very level of readability as compared to other programming languages.
- It is free and supports cross platform as python code can be called from any version of machine.
- Python has strong community support so if any problem is encountered, support is available online.
- Python has powerful tools available such as statsmodels, matplotlib, Pandas,Numpy and SciPyfor calculation intense projects

Since we have exhaustively used the `get_data_yahoo` function from the `pandas_datareader` python library we will briefly discuss the parameters it takes. Please note we utilized only those arguments which are relevant to the project requirements. From [10] parameter list as listed below:

- `symbols` : string, array-like object (list, tuple, Series), or DataFrame Single stock symbol (ticker), array-like object of symbols or DataFrame with index containing stock symbols.
- `start` : string, (defaults to '1/1/2010') Starting date, timestamp. Parses many different kind of date representations (e.g., 'JAN-01-2010', '1/1/10', 'Jan, 1, 1980')
- `end` : string, (defaults to today) Ending date, timestamp. Same format as starting date.
- `retry_count` : int, default 3 Number of times to retry query request.
- `pause` : int, default 0 Time, in seconds, to pause between consecutive queries of chunks. If single value given for symbol, represents the pause between retries.
- `session` : Session, default None requests.Session instance to be used
- `adjust_price`: bool, default False If True, adjusts all prices in hist data ('Open','High', 'Low','Close') based on 'Adj Close' price. Adds 'Adj Ratio' column and drops 'Adj Close'.
- `ret_index` : bool, default False If True, includes a simple return index 'Ret Index' in hist data.
- `chunksize` : int, default 25 Number of symbols to download consecutively before initiating pause.
- `interval` : string, default 'd' Time interval code, valid values are 'd' for daily, 'w' for weekly, 'm' for monthly and 'v' for dividend.

In our analysis, for the symbol parameter we are passing ticker symbol one at a time. Though, we have an option to pass multiple tickers as an array argument. We are using `get_data_yahoo` function and utilizing only first three parameters: symbols, start and end. This function returns `YahooDailyReader` object which can further be manipulated to get Open, High, Low, Close, Adj Close and Volume stock values. Since default number of retry count is three we will be using this default value. Default value of pause which is zero is also good with respect to our requirement so we will not pass this as argument. Session argument should be used when we are handling multiple request in parallel in the code since our project we just need one session so we will not use this argument. `adjust_price` is not required in our analysis since we are interested only in returns which can be fetched using `pct_change()` function. Since return index is of no use in calculating the returns, we will not use this argument. Argument `chunksize` is used to modify number of consecutive downloads of stocks since we are just using single ticker so this argument is of no use. This function uses interval also as a parameter since we are only interested in daily values and daily value is the default interval so we didn't pass this argument in the function call. We could also use the contemporary google function which is `get_data_google`. Arguments which goes to the `get_data_google` are symbols, start, end, `retry_count`, pause, `chunksize` and session since we are not using `get_data_google` function in our project we will not discuss these in detail.

## 7 INTRODUCTION TO CAPM MODEL

CAPM [Capital asset pricing model] model was developed by William Sharpe and John Lintner in 1964. This model is considered so powerful that it is being used in current prediction models. There are few advantages of using CAPM model as compared to other pricing models:

- This model is a single dimensional model and easy to use, still powerful to model capital asset pricing.
- Since this model is based on the market portfolio and risk free rate, this model removes unsystematic risk.
- We can run root mean square algorithm to validate the algorithm.
- This model provides a flexibility to utilize various risk free rates and run model for various time range.
- This model can be applied to various financial objects such as stocks, put option, call option, bonds, and etc

This model can be used to evaluate the theoretical expected return on a security, security can be any financial object such as stocks, put option, call option, bonds, and etc. In CAPM model we evaluate how much financial object is sensitive to the market using statistical analysis. Then this sensitivity which is also known as beta is used to find the expected return on security. This expected return can be on daily basis, weekly basis, monthly basis or yearly. Here is the formula to evaluate expected return:

$$E(R_i) = r_f + \beta_i(E(r_m) - r_f)$$

Where

- $E(R_i)$  is expected return
- $r_f$  is risk free interest rate example: Government bond
- $E(r_m)$  is return on market example SP 500

- $\beta_i$  is sensitivity of stock with respect to market

$\beta_i$  can further be defined how much stock is sensitive to the stock market. Example if  $\beta_i$  for a particular stock is two it means if market goes up by five percent then stock will go up by ten percent and if market goes down by two percent then stock will go down by ten percent. In terms of statistics  $\beta_i$  is defined as:

$$\beta_i = \frac{Cov(R_i, r_m)}{Var(r_m)}$$

Where covariance and variance are defined as

$$\bullet cov_{x,y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$\bullet var^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

$\beta_i$  matrix can be used to illustrate  $\beta_i$  in a following way:

	$\beta_i$	MarketReturn	ExpectetReturn
Row1	+2	+5%	+10%
Row2	-2	+5%	-10%
Row3	+0.5	+4%	+2%
Row4	+0.5	-4%	-2%

Above matrix suggests how expected returns can be correlated with the the  $\beta_i$ . Example if for certain company has  $\beta_i$  of +2 and market returns is +5 % then company's expected returns can be predicted as +10 %. Please note  $\beta_i$  can be positive as well as negative.

## 8 PROPOSED ANALYSIS

In this project we will utilize structured data and use CAPM [Capital asset pricing model] to statistically find the expected daily return of selected technological stocks: Amazon and Yahoo. This daily expected return can be used to predict next day stock value given the condition we have current stock price. Following formula can be used to predict next day stock price:

---

**Next Day stock price** =: Today stock price \* (1+Daily expected return)

---

Daily expected return will be calculated using CAPM model. Daily expected return sensitivity in CAPM terminology is also known as beta. In this project beta will be calculated based on two time frames:

---

**Time frame 1:** [01/01/2000 to 12/31/2016] 16 years of data

**Time frame 2:** [01/01/2006 to 12/31/2016] 10 years of data

---

Thus we we will have 2  $\beta_i$ :

$$\beta_1 = \frac{Cov(R_1, r_{m1})}{Var(r_{m1})}$$

$$\beta_2 = \frac{Cov(R_2, r_{m2})}{Var(r_{m2})}$$

Where

- $\beta_1$  is  $\beta$  based on time frame 1

- $\beta_2$  is  $\beta$  based on time frame 2
- $R_1$  is actual return based on time frame 1
- $r_{m1}$  is a mean market return based on time frame 1
- $R_2$  is actual return based on time frame 2
- $r_{m2}$  is a mean market return based on time frame 2

Above two time frames will be our training data set. We will run two analysis: one on training time frame 1 and other on training time frame 2 to arrive at predicted CAPM variables. Then we will use this training data set to predict stock returns for test data set which will comprise of time frame:

---

**Test data time frame:** 01/01/2017 to 11/16/2017

---

Then we will run the statistically analysis on the test data to evaluate if 16 years of training data produced more accurate result or else it added noise compared to 10 years of training data. Please note this is purely a quantitative analysis not qualitative. Actual returns can also be impacted by a qualitative factors such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc.

## 9 PROPOSED ALGORITHM

Code is written purely in python language and used the powerful rich python libraries such as statsmodels, matplotlib, Pandas, Numpy and SciPyfor. We have used jupyter notebook as interpreter tool to python. Code is started by importing above mentioned rich python libraries. Since we are interested only in technological stocks: Amazon and yahoo we need to initialize they stock ticker with the python variable. In CAPM model we need to know the market return in order to know the stock sensitivity we will also initialize market ticker with SP 500 index. As discussed above we will be using get\_data\_yahoo function from the pandas\_datareader and in this project we will be only utilizing only first three parameters which is stock ticker, start date and end date. For first iteration we will be using get\_data\_yahoo to fetch stocks and market returns for time frame 1. For having better understanding of how the data looks when fetched using get\_data\_yahoo function, we will have amazon financial data matrix calculated like:

```
amazonData = dr.get_data_yahoo('AMZN', start_date, end_date)
```

Where

- dr is pandas\_datareader.data class
- amazon is stock ticker for amazon which is 'AMZN'
- start\_date is start date of time frame 1: 01/01/2000
- end\_date is end date of time frame 1: 12/31/2016

and synopsis of above amazon data looks like:

	Open	High	Low	Close	Adj	Volume
23 - Dec - 16	764.54	766.50	757.98	760.59	760.59	1976900
27 - Dec - 16	763.40	774.65	761.20	771.40	771.40	2638700
28 - Dec - 16	776.25	780.00	770.50	772.13	772.13	3301000
29 - Dec - 16	772.40	773.40	760.84	765.15	765.15	3153500
30 - Dec - 16	766.46	767.40	748.28	749.86	749.86	4139400

Similarly using get\_data\_yahoo we will fetch Yahoo and market returns. Since we are interested in daily return, we fetched the daily data from yahoo finance which is evident from the above result data set. Now lets find the percentage change on the daily Close value to get the percentage change array which in finance terminology will be daily return on stock. For finding the percentage change we are using pct\_change() function on the close column of result set. This function can be elaborated as follows:

```
return_amazon = amazonData.Close.pct_change()[1 :]
return_yahoo = yahooData.Close.pct_change()[1 :]
return_market = marketData.Close.pct_change()[1 :]
```

return\_amazon, return\_yahoo and return\_market are two dimensional arrays and we need to convert them to single dimensional array in order to run statistical analysis. We can use dot values method to extract single dimensional array out of 2 dimensional array. This operation can be elaborated as follows:

```
X_amazon_actualReturns = return_amazon_testing.values
X2_yahoo_actualReturns = return_yahoo_testing.values
Y_market_actualReturns = return_market_testing.values
```

Please note these are actual returns - fetched from yahoo finance. Now in order to evaluate expected return for the testing period based on the calculated beta we need to calculate the risk free rate  $r_f$  as mentioned above in the CAPM formula. Please note get\_data\_yahoo formula will fetch the annualized rate but here we are dealing with the daily returns so this needs to be normalized to daily rate. Here we are using Treas Yld Index-10 Yr Nts bond. Ticker symbol for Treas Yld Index-10 Yr Nts bond is TNX. Please note get\_data\_yahoo will return columns: Open, High, Low, Close, Adj Close and Volume. Dot values will convert to 2 dimensional array and then used index [0][4] to fetch annual rate. Detailed code with comments is mentioned on jupyter notebook.

Conversion of annualized return to daily return can be done using following formula:

```
riskFreeDailyRate = (1 + riskFreeAnnualRate)(1/365) - 1
```

Now we need to copy the content of X\_amazon\_actualReturns to new array X\_amazon\_predictedReturns and initialized each element in X\_amazon\_predictedReturns using CAPM model as discussed above in Introduction:

```
X_amazon_predictedReturns = list(X_amazon_actualReturns)
```

We will do the same for Yahoo stocks:

```
X2_yahoo_predictedReturns = list(X2_yahoo_actualReturns)
```

In the code we have run the while loop and each element of X\_amazon\_predictedReturns and X2\_yahoo\_predictedReturns is assigned the value based on CAPM model. Now we have two returns arrays for amazon stocks based on sixteen years of data:

- X\_amazon\_actualReturns are the actual returns
- X\_amazon\_predictedReturns are returns based on the CAPM model.

Similarly we have two returns arrays for yahoo stocks based on sixteen years of data:

- X2\_yahoo\_actualReturns are the actual returns
- X2\_yahoo\_predictedReturns are the returns based on the CAPM model.

Now we can utilize mean\_squared\_error function from the sklearn.metrics python library to find how predicted returns are deviated from the actual returns. We will run this function on both stocks, amazon and yahoo:

```
a1 = Y_market_actualReturns
a2 = X_amazon_predictedReturns
y1 = Y_market_actualReturns
y2 = X2_yahoo_predictedReturns

rms_amazon = sqrt(mean_squared_error(a1,a2))
rms_yahoo = sqrt(mean_squared_error(y1,y2))
```

Here is the root mean square values for both the stocks under sixteen years of data case:

- Root mean square error for Amazon stocks analysis based on 16 years of data 0.0013770 or 0.137 percent
- Root mean square error for Yahoo stocks analysis based on 16 years of data 0.0014313 or 0.143 percent

Now we run the same analysis as discussed above for the ten years of data and will validate how much predicted stocks returns based on the ten years of data are deviated from the actual returns using root mean square method. Please note testing data set remains the same we are just using different training data set. This will let us compare if sixteen years of data is of more worth in predicting stock returns or it added noise to the analysis:

- Root mean square error for Amazon stocks analysis based on 10 years of data 0.0005310 or 0.053 percent
- Root mean square error for Yahoo stocks analysis based on 10 years of data 0.0014910 or 0.149 percent

Above analysis is purely quantitative and does not include any elements of qualitative analysis. It shows predicting yahoo stock price or its returns based on the sixteen years of data or ten years of data - both resulted in almost same results. However things are totally different for the amazon stocks, recent ten years of amazon stocks data produced more accurate results as compared to using recent sixteen years of data. Author of [7] agrees with the fact that most recent financial data are the better predictors of the future price returns. Though, in the [7] author has used the neural networks and support vector machine for prediction. Author also stressed that neural networks algorithm produced better accuracy than other machine learning algorithms.

## 10 THREE PARADIGMS OF PREDICTION

Data prediction and analysis done in this project is purely quantitative. However there are other paradigms of predictions also which we will discuss here. Example in above analysis we totally missed the qualitative aspect of the data. This is why it explains recent data on amazon stocks produced better results. Here are the other prediction paradigms explained by the author of [3] :

- Quantitative research based prediction. This is a method where we utilize statistical tools to arrive at predictive value based on data
- Quantitative research based prediction. This is a method where we utilize conceptual knowledge to arrive at predicted value. Example of such study would be prediction of stocks based on the events such as mergers acquisitions, bankruptcy, Fraud, political changes, market crashes, housing bubble, dot net bubble and etc.
- Mixed research based prediction. This is a hybrid method where we utilize both qualitative and quantitative results to predict result.

In this project we used quantitative based approach to validate the fact if more data is good for prediction or it adds noise. Result of this project also showed there is a importance of recent data in predicting results. This is also validated by the research done under [7]

## 11 FUTURE WORK

In this project, analysis is based on two technological stocks: yahoo and amazon. We can extend our study to more diverse portfolio by including more stocks from various industries. Technological stocks tend to be more volatile than other stocks. Since this project is purely quantitative based prediction we deliberately choosen the technological stocks to leverage their volatility. More accurate prediction could also be made by encapsulating qualitative based prediction in the analysis which is more like a hybrid approach. Such hybrid approaches includes assigning weight to each predictions and taking cumulative result. As the part of future work we can also compare results across industry and arrive at conclusion which industry is more stable in prediction. Comparison can based on the root mean square analysis which is discussed in this project report.

## 12 CONCLUSION

Main objective of doing this project is to know the importance of big data in predicting financial variables. Analysis of this project is based on two stocks: amazon and yahoo. We started this project report with the discussion of importance of big data in financial industry. In the introduction we discussed how various industries are investing in the Big Data to attain higher standards in terms of quality and customer satisfaction. Then we discussed what are the various types of data available: structured and unstructured. Since in this project we utilized only structured data so it was discussed deeply. This project report also touch base with various challenges financial industry takes in utilizing the value of big data. As the part of literature review we reviewed various researches done in the field of stock returns prediction. In the financial data extraction section we reviewed various technical requirements need

for financial data extraction. As the part of data analysis for this project we discussed what are the various ways to fetch live stock data from yahoo or google server. In this project we used the rich financial python libraries for the analysis so we discussed them in details in this report. Financial model which we chose for the prediction is the CAPM model which is explained theoretically in this report. There are two different section where we discussed the proposed analysis and proposed algorithm. We finally concluded the report by discussing three paradigms of prediction. In the end we also mentioned what further can be done under future work section.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and TAs for their support and suggestions to write this paper. TAs and professor are very good in terms of providing valuable guidance and suggestion in a very prompt fashion.

## REFERENCES

- [1] Qasem Al-Radaideh, Adel Abu Assaf, and Eman Alnagi. 2013. Predicting Stock Prices Using Data Mining Techniques. (12 2013). [https://www.researchgate.net/publication/281865047\\_Predicting\\_Stock\\_Prices\\_Using\\_Data\\_Mining\\_Techniques](https://www.researchgate.net/publication/281865047_Predicting_Stock_Prices_Using_Data_Mining_Techniques)
- [2] A. F. Atiya. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12, 4 (Jul 2001), 929–935. <https://doi.org/10.1109/72.935101>
- [3] Adam Chu. 2017. Quantitative, Qualitative, and Mixed Research. (2017). <https://www.bcps.org/offices/lis/researchcourse/images/lec2.pdf>
- [4] Daniel D. Gutierrez. 2014. *Big Data for Finance*. Technical Report. Dell & Intel. [https://whitepapers.em360tech.com/wp-content/files\\_mf/1427803213insideBIGDATAGuidetoBigDataforFinance.pdf](https://whitepapers.em360tech.com/wp-content/files_mf/1427803213insideBIGDATAGuidetoBigDataforFinance.pdf)
- [5] Kazim Hussain and Elsa Prieto. 2015. *Big Data in Finance*. Chapman and Hall/CRC, <https://www.cs.helsinki.fi/u/jilu/paper/bigdataapplication04.pdf>, Chapter 17, 329–356.
- [6] Kazim Hussain and Elsa Prieto. 2016. *Big Data in the Finance and Insurance Sectors*. Springer, Cham, "<https://link.springer.com/content/pdf/10.1007/>", Chapter 12, 209–223.
- [7] Hui Lin. 2014. Stanford. (2014). <https://pdfs.semanticscholar.org/56f0/59ea400f31b60bfde4d59aea71bd7b411553.pdf>
- [8] Burton G. Malkiel. 2015. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. Recorded Books on Brilliance Audio. <https://www.amazon.com/Random-Walk-Down-Wall-Street/dp/1501260375?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1501260375>
- [9] A. Parsian, W. Yeoh, and M. S. Ee. 2015. Quality-Based SQL: Specifying Information Quality in Relational Database Queries. *Computer* 48, 9 (Sept 2015), 69–74. <https://doi.org/10.1109/MC.2015.264>
- [10] Kevin Sheppard. 2017. daily.py. (2017). [https://github.com/pydata/pandas-datareader/blob/master/pandas\\_datareader/yahoo/daily.py](https://github.com/pydata/pandas-datareader/blob/master/pandas_datareader/yahoo/daily.py)
- [11] Philip M. Tsang, Paul Kwok, S.O. Choy, Reggie Kwan, S.C. Ng, Jacky Mak, Jonathan Tsang, Kai Koong, and Tak-Lam Wong. 2007. Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence* 20, 4 (2007), 453 – 461. <https://doi.org/10.1016/j.engappai.2006.10.002>
- [12] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. 2014. Modern modelling techniques are data hungry a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 14, 1 (22 Dec 2014), 137. <https://doi.org/10.1186/1471-2288-14-137>
- [13] Martin Walker and Mamoun Al-Debi'e. 2000. Fundamental Information Analysis: An Extension and UK Evidence. *ACS Biomaterials Science & Engineering* 31 (02 2000).
- [14] Mu-Cherng Wu, Sheng-Yu Lin, and Chia-Hsin Lin. 2006. An effective application of decision tree to stock trading. *Science Direct* 31 (08 2006), 270–274.
- [15] Sonja Zillner, Tilman Becker, and Munn. 2016. *Big Data-Driven Innovation in Industrial Sectors*. Springer International Publishing, Cham, Chapter 4, 169–178. [https://doi.org/10.1007/978-3-319-21569-3\\_9](https://doi.org/10.1007/978-3-319-21569-3_9)

## A HID 301:GAGAN ARORA

- Identified Project topic.
- Collected the python financial libraries.
- fetched data from yahoo finance
- Studied, designed and reviewed CAPM model
- Implemented CAPM model using python libraries
- Created 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/hid301/blob/master/project/finalProject.ipynb>

```
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 "Ref7" isn't style-file defined
--line 169 of file report.bib
Warning--entry type for "Ref8" isn't style-file defined
--line 187 of file report.bib
Warning--entry type for "Ref14" isn't style-file defined
--line 341 of file report.bib
Warning--entry type for "Ref15" isn't style-file defined
--line 359 of file report.bib
Warning--empty address in Ref13
Warning--page numbers missing in both pages and numpages fields in Ref10
(There were 6 warnings)
```

```
bibtext _ label error
```

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

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

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

```
=====
[2017-12-10 13.50.12] 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 1.1s.
./README.yml
25:24     error    trailing spaces  (trailing-spaces)
33:4      error    wrong indentation: expected 4 but found 3  (indentation)
33:17     error    trailing spaces  (trailing-spaces)
35:4      error    wrong indentation: expected 7 but found 3  (indentation)
37:4      error    wrong indentation: expected 7 but found 3  (indentation)
```

```
39:56      error      trailing spaces  (trailing-spaces)
51:25      error      trailing spaces  (trailing-spaces)
```

---

Compliance Report

---

```
name: Arora, Gagan
hid: 301
paper1: 100% Oct 29 17
paper2: 100% Nov 4
project: 100% 12/2/2017
```

```
yamlcheck
```

---

```
wordcount
```

---

```
7
wc 301 project 7 5857 content.tex
wc 301 project 7 5970 report.pdf
wc 301 project 7 866 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
```

---

```
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--entry type for "Ref7" isn't style-file defined
--line 169 of file report.bib
Warning--entry type for "Ref8" isn't style-file defined
--line 187 of file report.bib
Warning--entry type for "Ref14" isn't style-file defined
--line 341 of file report.bib
Warning--entry type for "Ref15" isn't style-file defined
--line 359 of file report.bib
Warning--empty address in Ref13
Warning--page numbers missing in both pages and numpages fields in Ref10
(There were 6 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

---

312: \textbf{\textit{Next Day stock price }}: Today stock price \*
(1+Daily expected return)\newline

320: \textbf{\textit{Time frame 1}}: [01/01/2000 to 12/31/2016] 16
years of data\newline

321: \textbf{\textit{Time frame 2}}: [01/01/2006 to 12/31/2016] 10  
years of data\nline

350: \textbf{\textit{Test data time frame}}: 01/01/2017 to  
11/16/2017\nline

passed: False

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

find cite {

---

passed: True

# Predicting Housing Prices - Kaggle Competition

Murali Cheruvu, Anand Sriramulu

Indiana University

3209 E 10th St

Bloomington, Indiana 47408

mcheruvu@iu.edu, asriram@iu.edu

## ABSTRACT

Apply exploratory data analysis and implement various advanced supervised machine learning algorithms to predict neighborhood housing sale prices found in the sample test dataset. Compare the predicted models and results from these advanced supervised algorithms. Apply ensembled model to achieve better predictions, hence get good score in kaggle competition.

## KEYWORDS

i523, hid306, Supervised Learning Algorithms, Exploratory Data Analysis, Kaggle

## 1 BACKGROUND

Kaggle website hosts a platform for data science competitions. Predicting housing prices is a competition project maintained by Kaggle. This project has two data sets - *train* and *test*, with 79 exploratory variables describing almost every aspect of residential homes in city of Ames, Iowa state. The goal of this competition is to predict the sale prices of residential homes listed in the test dataset as accurately as possible. Submissions are evaluated on *root mean squared error* (RMSE) between the logarithm of predicted value and the logarithm of observed sale price, also known as *kaggle score*. Submitted file expects two variables: row id and the predicted sale price of each home listed in the test dataset. Our goal is to participate in this competition and get placed within top 20% of the submitted algorithms.

## 2 INTRODUCTION

Training dataset of the *predicting housing sale prices* project contains sale price of the homes, and using this training data set, how accurately we can predict sale prices of the homes in the test dataset using preprocessing and thorough data analysis. Many developers used advanced learning algorithms - XGBoost, Lasso and Neural Network, to predict the sale prices in the competition and achieved better kaggle scores. We have applied various exploratory analysis techniques and engineer the features before applying a few advanced supervised learning algorithms to create more accurately predicted models.

## 3 DOMAIN KNOWLEDGE

To predict accurate *sale price*, we will need to understand the domain well. We need to build the intuition around all the exploratory variables in the dataset and focus on which factors could influence the target variable: *sale price*. If we do not find all these factors, perhaps, we need to add new features to address the gaps in dataset describing the domain. Some of the factors which, we think, can directly influence house prices are:

- What is the overall Size or area of the house?
- How good is the location of the house - closer to highways?
- How good is the neighborhood?
- How old is the house?
- What is the quality of the construction?
- How many garages are there in the house?
- What are the floor plans?
- How many number of bedrooms are there in the house?
- How many number of bathrooms are there in the house?
- What is the size of living area?

## 4 EXPLORATORY DATA ANALYSIS

We can start the process with exploratory data analysis. There are 1460 rows in the training data set and 1459 rows in the test dataset. Out of the 80 variables, 23 are nominal, 23 are ordinal, 14 are discrete, and 20 are continuous. We have combined training and testing datasets for easier analysis. We excluded *Id* attribute as it does not add value in the modeling. We also removed *Sale Price*, the target variable, from the training dataset. All the variables are listed in the appendix section as a reference.

### 4.1 Handling Missing Values

First part of the analysis was to check for any missing values in the training and testing datasets as shown in figure (1). Using the bar plot shown in figure (2), we have identified that there are 5 variables: *pool quality*, *miscellaneous features*, *alley*, *fence* and *fire place quality*, having the most missing data. All the missed values for the numeric variables are analyzed further to decide whether we need to delete the instances of all the data with missing values or impute them using meaningful data such as median of the corresponding variable.

[Figure 1 about here.]

[Figure 2 about here.]

### 4.2 Analyze Numerical Variables

There are 37 numerical variables after excluding the *Id* variable. We have analyzed all the numerical variables for data patterns such as skewness in the data and range of the possible values. We have shown *sale price*, *overall quality*, *garage live area* and *year built*, in the figures (4) and (5) as a few sample plots from the numerical analysis. Corresponding code snippet is shown in figure (3).

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

### 4.3 Analyze Categorical Variables

There are 43 categorical variables in the combined dataset. We have analyzed all categorical variables and found the ways to fill the missing values. We have also evaluated proper approaches to convert them into numerical factors. Later on in the feature engineering section, we will go through more details on numerical factors. Categorical variable factors and the corresponding code snippet for *neighborhood* and *sale type* are shown in figure (6) and figures (7).

[Figure 6 about here.]

[Figure 7 about here.]

### 4.4 Analyze Correlations

*Numpy* package offers correlations functionality to analyze the variables that are positively or negatively correlated with the *sale price* and also analyze any interdependencies among the variables. Figure (8) and (9) shows the code snippet and the correlations plot. From that we can list the top 10 features those are strongly correlated with the target variable - *sale price*. We can visualize a few pair-wise correlation graphs with *sale price* for further detailed analysis. Figures (10) and (11) show how *overall quality*, *ground live area*, *garage cars* and *garage area* are positively correlated with *sale price*.

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

- (1) OverallQual: Overall material and finish quality
- (2) GrLivArea: Above ground living area square feet
- (3) GarageCars: Size of garage in car capacity
- (4) GarageArea: Size of garage in square feet
- (5) TotalBsmtSF: Total square feet of basement area
- (6) 1stFlrSF: First Floor square feet
- (7) FullBath: Full bathrooms above grade
- (8) TotRmsAbvGrd: Total rooms above ground
- (9) YearBuilt: Original construction date
- (10) GarageYrBlt: Garage built year

### 4.5 Skewed Data Analysis

From the numerical analysis, we have identified that there are a few numerical variables need further analysis to identify the skewed data. We did not find any key variables those have skewed more than 75%. However, we wanted to replace the *sale price* with corresponding logarithmic value for the predictive models and later convert it back to the exponential value before submitting to the kaggle competition. Figure (12) shows the *sale price*, before and after applying the logarithmic value.

[Figure 12 about here.]

### 4.6 Outlier Analysis

Continuing with exploratory analysis, we have analyzed the outliers using *Cook's distance*. Cook's distance is a measure calculated from a regression model to find out the influence exerted by each

observation (row) on the predictions. As a practice, those observations that have a Cook's distance greater than 4 times the mean value may be classified as an outlier. Outlier detection can be done using univariate and multivariate analysis. In univariate model, the outliers are those observations that are present outside of  $1.5 * \text{IQR}$ , where IQR (*Inter Quartile Range*) is the difference between 75th and 25th quartiles. Analyzing outliers in any observations based on single variable may lead to incorrect inferences. Cook's distance generalizes the outlier analysis using multivariate approach [5]. Figure (13) is the code implementing Cook's distance to find the outliers from training dataset and figure (14) shows the scatter plot with outliers being marked as bubbles. The bigger the bubbles, the bigger outlier deviations from the mean value. We have further analyzed two key variables - *ground live area* and *garage area* that are in high correlation with the *sale price*. From the scatter plot shown in figure (15), we can see that *garage live area* has 4 outliers with values greater than 4,000 sq ft. We can also visualize 4 outliers in *garage area* scatter plot with values greater than 1,200 sq ft. as shown in figure (16). We have removed the 8 outlier rows related to these two variables from the training dataset, the corresponding code snippet shown in figure (17).

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

### 4.7 Feature Engineering

Feature engineering is a technique to analyze all the variables those influence target variable for better predictions. Part of feature engineering, we may need to create new features to make the data to be more expressive. One of the key intents, in analyzing categorical variables, is to convert them into numerical factors as most of the machine learning algorithms expect all the variables to be numeric for them to work more effectively. Feature engineering is a difficult task; majority of the effort is manual and requires lots of domain knowledge.

**4.7.1 Numerical Encoding.** Some of the categorical variables are ordinal. we can use T-shirt sizes: small, medium and large as an example to explain an ordinal variable. When we convert this category variable into numeric encoding, we need to retain the fact that there is an implicit order within the values. Supposing, we give ordinal encoding as - small = 1, medium = 2 and large = 3; we will satisfy the implicit order or weightage and that helps in modeling the system by elevating the importance of this implicit ordering in the values of the ordinal variable. There are a few other encoding techniques, such as one-hot, binary, polynomial and helmert to factorize categorical variables. We will use ordinal and one-hot encoding techniques for this dataset. Following are a few categorical variables converted to numerical:

- *Lot shape* is encoded as: 1 - regular, 2 - Irregular-I, 3 - Irregular-II, 4 - Irregular-III
- *Alley* is encoded as: 1 - none, 2 - gravel, 3 - paved
- All quality variables such as *garage quality* are encoded as: 0 - none, 1 - poor, 2 - fair 3 - typical 4 - good, 5 - excellent

- *Building type* is encoded as: 1 - single-family, 2 - two-family, 3 - duplex, 4 - townhouse end unit, 5 - townhouse inside unit
- *Overall quality* is encoded as: 1 to 3 - bad, 4 to 6 - average, 7 to 10 - good

**4.7.2 One-hot Encoding.** One-hot encoding converts the category variable into many binary vectors, one new numeric variable for each value in the category. Assume that we have a categorical variable called signal-light with three possible values: green, yellow and red. We will need to convert these values into numeric - green = 1, yellow = 2 and red = 3. When we apply one-hot encoding on this variable, basically we are creating three new categorical variables - signal-light-green, signal-light-yellow and signal-light-red along with the original variable - signal-light, each is pretty much a binary vector having 1s for all the corresponding values; otherwise 0s. With hot-encoding, we are basically increasing dimensions in the model. After extensive feature engineering applied on the housing dataset, we have added 228 new features (variables). Figure (18) shows the python methods to factorize categorical variables and one-hot encoding techniques.

[Figure 18 about here.]

**4.7.3 New Features.** By adding new features that fill the gaps in domain model, we can guide the model predictions more accurately. We can, easily, create more meaningful new features from existing features, such as:

- What is the total area of the house? - This variable is sum of 18 existing variables that are contributing to the overall size of the house, such as *lot frontage*, *lot area*, *ground live area*, *pool area* and *garage area*.
- Whether house has been ever remodeled? - We can find this out using two variables: *year built* and *year remodel added*.
- House remodeled since? - We can find this out using two variables: *year sold* and *year remodel added*.
- Is it a very new house? - This can be calculated based on *year built*
- What is the age of the house? - This is a calculated value from *year build* (formula: 2010 - *year built*)
- When was it last sold? - This is a calculated value from *year build* (formula: 2010 - *year sold*)
- Which season house was last sold in? - This is a calculated value from *month build*

## 5 ALGORITHMS AND METHODOLOGY

Linear regression predicts the target variable using best possible straight line fit to the set predictor variables. The best fit is usually the one that minimizes the root mean squared error (RMSE) between the actual and predicted data points. However, with complex problem space such as the housing prices dataset, we have lots of variables relating to the target variable in a non-linear fashion. Trivial supervised learning algorithms will not be effective to provide accurate *sale price* predictions. To overcome this challenge, we have applied various advanced supervised learning algorithms, such as Support Vector Machine (SVM), Random Forest, Lasso, Ridge, XG-Boost and Neural Network, to predict the test data housing prices.

### 5.1 Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) algorithms can be used to solve classification and regression problems. SVM regression relies on kernel functions for modeling the data. SVM creates larger margins between categories of data so that they are linearly separable. SVM handles non-linearly separable data, mainly for regression problems, using kernel functions, such as polynomial, radial basis function (RBF) and sigmoid, to project the data onto a hyperplane. Figure (19) shows the python implementation for *sale price* predictions of the housing test dataset.

[Figure 19 about here.]

### 5.2 Random Forest Algorithm

Random Forest is an advanced machine learning algorithm for predictive analytics. Random Forest ensembles multiple decision trees to create an additive learning model from the sequence of base models created by each decision tree that worked on a sub-sample dataset. Random Forest models are suitable to handle tabular datasets with hundreds of numeric and categorical features. Along with missing values, non-linear relations between features and the target, will be handled well by random forest algorithms. With proper tuning of hyper-parameters of the random forest algorithm, it can perform well with decent accuracy in the predictions without overfitting the model. Unlike similar regression models, it does not offer feature coefficient information but it provides *feature ranking* functionality very nicely. Figure (20) shows the random forest algorithm details for the *sale price* predictions implemented using *sklearn* package and the figure (21) shows the top 10 important features selected by random forest to model the predictions.

[Figure 20 about here.]

[Figure 21 about here.]

### 5.3 Lasso Algorithm

Lasso is a regression model that uses shrinkage to bring data points towards the center, similar to the mean value of all the data points. Lasso stands for Least Absolute Shrinkage and Selection Operator. It is a regularized linear model with penalty term *lambda* to minimize the error. Parameter penalization controls overfitting the input data by shrinking variable coefficients to 0. Essentially this makes the variables no effect in the model, hence reduces the dimensions. Figure (22) shows the lasso algorithm implementation for *sale price* predictions in python.

[Figure 22 about here.]

### 5.4 Ridge Algorithm

Ridge algorithm is very similar to lasso algorithm with the same goal. While lasso performs *L1 regularization*, ridge applies *L2 regularization* techniques in modeling the predictions. L1 regularization adds penalty to the variables equivalent to *absolute value of the magnitude* of the coefficients, whereas L2 adds the penalty equivalent to *square of the magnitude* of the variable coefficients. Figure (23) shows the python implementation of the ridge algorithm for the *sale price* predictions. Figures (24) and (25) show the top 10 positively and top 10 negatively influencing variables with *sale price*.

[Figure 23 about here.]  
[Figure 24 about here.]  
[Figure 25 about here.]

## 5.5 XGB Boosting Algorithm

XGBoost (eXtreme Gradient Boosting) is one of the Gradient Boosted Machine algorithms. It ensembles (combines) optimized model by taking trained models from all the preceding iterations. XGBoost regularizes the variables (parameters) to reduce the overfit and can work well with variables having missing values. It is empowered with built-in cross validation to reduce the boosting iterations; hence offers better performance along with parallel processing on distributed systems such as Hadoop. By tuning the XGBoost hyper parameters, we can achieve well optimized model that can make more accurate predictions. XGBoost uses *F-Score* to measure the importance of variables. Table (1) explains the hyper-parameters of XGBoost algorithm and also given the python code, as shown in figure (26), implementing for *sale price* predictions. Figure (27) shows the top 10 feature selection by the XGBoost.

[Table 1 about here.]  
[Figure 26 about here.]  
[Figure 27 about here.]

## 5.6 Neural Network Algorithm

Neural Network is, a *directed graph*, organized by layers and layers are created by number of interconnected neurons (or nodes). Every neuron in a layer is connected with all the neurons from previous layer; there will be no interaction of neurons within a layer. The performance of a Neural Network is measured using *cost or error function* and the dependent input *weight* variables. *Forward-propagation* and *back-propagation* are two techniques, neural network uses repeatedly until all the input variables are adjusted or calibrated to predict accurate output. During, forward-propagation, information moves in forward direction and passes through all the layers by applying certain weights to the input parameters. *Back-propagation* method minimizes the error in the *weights* by applying an algorithm called *gradient descent* at each iteration step. We have used *TensorFlow* python library to predict the *sale price* of housing dataset using simple feed-forward neural network. TensorFlow uses *tensors*, special multi-dimensional arrays to store the datasets for easier linear algebra and vector calculus operations.

## 5.7 Model Ensembling

We can create a robust predictive model with better accuracy by merging two or more machine learning algorithms. This technique is called *model ensembling*. Ensembled algorithms may be similar in functionality or may entirely be different from each other. Individual algorithms may not perform great but by ensembling them, the overall system can offer much better performance and accuracy. Variations in the predicting logic in each of these individual algorithms will bring unbiasedness into the unified model. *Bagging*, *boosting* and *stacking* are popular ensembling techniques. Many of the advanced machine learning algorithms use ensembled approaches to achieve accurate classifications or predictions. Random Forest uses bagging, XGBoost uses boosting and Neural Network

applies stacking ensembling techniques. For the kaggle submission, we have created an ensembled model by averaging *Sale Price* of the top 3 performing ensembled algorithms - XGBoost, Lasso and Neural Network. As predicted, ensembled model has scored better compared to the individual algorithms. By applying advanced machine learning algorithms, we have placed our scores within top 20% of the competition. Table (2) displays each algorithm and the *root mean squared error* (RMSE) along with the *kaggle score*.

[Table 2 about here.]

# 6 DEVELOPMENT ENVIRONMENT

## 6.1 OS and Programming Language

We have used *Ubuntu 16.4* Operating System that runs in Windows 10 through Oracle Virtual Box 5.2. Python 2.7 has been used as the programming language for this project. Data visualizations are done using *seaborn* and *matplotlib* packages. Most of the algorithms implemented in this project are using *sklearn* package. For the neural network algorithm, we have used *tensorflow* package as it offers simple programming interface to the complex processing needed by the algorithm. Our code is placed in gitub repository at *git@github.com:bigdata-i523/hid306.git*.

## 6.2 Project Folder Structure

Project is organized in three folders - code, data and images. Code folder has all the python code files. Data folder contains the *house pricing* sample datasets provided by Kaggle website for the housing prices competition that we used for the exploratory analysis and *sale price* predictions. We also stored all the *sale price* prediction output files from various algorithms in the data folder. Images folder contains all the data visualization files that we have created during the analysis and in processing the algorithms. We wanted to create interactive and sharable code files that contain not only the python code but also corresponding explanation along with data visualizations. Jupyter Notebook application is ideal for such facilitation with python code components. Using Jupyter Notebook, it would be easy to share live code with the reviewers. Such environment allows to explore the code-base easily along with the interactive code execution and visualize all the corresponding exploratory analysis results with the graphs.

## 6.3 Project Files

We have a total of 11 Jupyter Notebook driven python code files. First 4 files are focused on doing the exploratory data analysis and the next 6 files are meant for six supervised machine learning algorithms - SVM, Random Forest, Ridge, Lasso, Neural Network and XGBoost. Last code file is dedicated for ensembling the top 3 algorithms with best predictions of housing sale prices in the test dataset. We have named them in a sequence as there is an implicit order in the execution of these files. We wanted to do the data analysis first before running the predictive algorithms.

## 6.4 List of Code Files

Following is the list of code files:

- (1) Code: 1.1\_exploratory\_analysis\_numerical.ipynb - To load datasets and analyze all numerical variables

- (2) Code: 1.2\_exploratory\_analysis\_categorical.ipynb - To analyze categorical variables in the dataset
- (3) Code: 1.3\_outlier\_and\_skewed\_data\_analysis.ipynb - Handles outlier and skewed data analysis
- (4) Code: 1.4\_feature\_engineering.ipynb - All the feature engineering is done in this file
- (5) Code: 2.1\_algorithm\_svm.ipynb - Implementation of SVM algorithm
- (6) Code: 2.2\_algorithm\_random\_forest.ipynb - Implementation of Random Forest algorithm
- (7) Code: 2.3\_algorithm\_ridge.ipynb - Implementation of Ridge algorithm
- (8) Code: 2.4\_algorithm\_lasso.ipynb - Implementation of Lasso algorithm
- (9) Code: 2.5\_algorithm\_neural\_network\_tf.ipynb - Implementation of Neural Network algorithm
- (10) Code: 2.6\_algorithm\_xgboost.ipynb - Implementation of XGBoost algorithm
- (11) Code: 3\_ensemble\_kaggle\_submission.ipynb - Implementation of Ensembled algorithm

## 6.5 List of Data Files

Following is the list of data files:

- (1) Input data file: train.csv - Sample training dataset with housing attributes along with the sale price
- (2) input data file: test.csv - Sample testing dataset similar to training dataset without the sale price
- (3) Output data file: kaggle\_python\_svm.csv - Predicted Housing Sale Prices from SVM algorithm
- (4) Output data file: kaggle\_python\_random\_forest.csv - Predicted Housing Sale Prices from Random Forest algorithm
- (5) Output data file: kaggle\_python\_ridge.csv - Predicted Housing Sale Prices from Ridge algorithm
- (6) Output data file: kaggle\_python\_xgboost.csv - Predicted Housing Sale Prices from XGBoost algorithm
- (7) Output data file: kaggle\_python\_lasso.csv - Predicted Housing Sale Prices from Lasso algorithm
- (8) Output data file: kaggle\_python\_neural\_network.csv - Predicted Housing Sale Prices from Neural Network algorithm
- (9) Output data file: kaggle\_python\_ensemble.csv - Predicted Housing Sale Prices from Ensembled algorithm

## 7 CONCLUSION

Generally, ensemble models performs better compared to individual algorithms. However, there are a few factors that influence accuracy and performance of the algorithms, such as handcrafted feature engineering, proper cost function with regularized input to address non-linearities in the training datasets and tuning hyper-parameters of the algorithms. While Deep Learning Neural Networks are good for image processing, K-Nearest Neighbor algorithms can handle unsupervised datasets with less complexity. Domain knowledge and algorithm selection play vital role in getting accurate predictions. XGBoost, Random Forest, Lasso and Neural Networks are advanced machine learning algorithms dominating in the data science competitions for classification and regression related tasks. With ensembling and iterative learning techniques, they can scale

well and offer better predictions for huge datasets having large number of features.

## A SAMPLE DATASET FILE DETAILS

The training and testing sample datasets provided by Kaggle contain the same variables explaining the housing real estate aspects. Training dataset contains the sale price information whereas the testing dataset does not the sale price as that is the target variable we need to predict using supervised machine learning algorithm. Following are the list of variables describing the housing real estate domain. Good understanding of the domain is needed for better exploratory data analysis and to apply the matching machine learning algorithms to the problem space.

- Id: Row Id
- SalePrice: Sale price of the house in dollars. This is the target variable to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area

- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: Dollar Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

## A.1 Factorization of categorical variables

Following are the factorized categorical variable details:

### A.1.1 Street (Nominal). : Type of road access to property

- Grvl - Gravel
- Pave - Paved

### A.1.2 Alley (Nominal). : Type of alley access to property

- Grvl- Gravel
- Pave - Paved
- NA - No alley access

### A.1.3 Lot Shape (Ordinal). : General shape of property

- Reg - Regular
- IR1 - Slightly irregular
- IR2 - Moderately Irregular
- IR3 - Irregular

### A.1.4 Land Contour (Nominal). : Flatness of the property

- Lvl - Near Flat /Level
- Bnk - Banked - Quick and significant rise from street grade to building
- HLS - Hillside - Significant slope from side to side
- Low - Depression

### A.1.5 Land Slope (Ordinal). : Slope of property

- Gtl - Gentle slope
- Mod - Moderate Slope
- Sev - Severe Slope

### A.1.6 Utilities (Ordinal). : Type of utilities available

- AllPub - All public Utilities (E,G,W, and S)
- NoSewr - Electricity, Gas, and Water (Septic Tank)
- NoSeWa - Electricity and Gas Only
- ELO - Electricity only

### A.1.7 Lot Config (Nominal). : Lot configuration

- Inside - Inside lot
- Corner - Corner lot
- CulDSac - Cul-de-sac
- FR2 - Frontage on 2 sides of property
- FR3 - Frontage on 3 sides of property

### A.1.8 Neighborhood (Nominal). : Physical locations within Ames city limits (map available)

- Blmngtn - Bloomington Heights
- Blueste - Bluestem
- BrDale - Briardale
- BrkSide - Brookside
- ClearCr - Clear Creek
- CollgCr - College Creek
- Crawfor - Crawford
- Edwards - Edwards
- Gilbert - Gilbert
- Greens - Greens
- GrnHill - Green Hills
- IDOTRR - Iowa DOT and Rail Road
- Landmrk - Landmark
- MeadowV - Meadow Village
- Mitchel - Mitchell
- Names - North Ames
- NoRidge - Northridge
- NPkVill - Northpark Villa
- NridgHt - Northridge Heights
- NWAmes - Northwest Ames
- OldTown - Old Town
- SWISU - South and West of Iowa State University
- Sawyer - Sawyer
- SawyerW - Sawyer West
- Somerst - Somerset
- StoneBr - Stone Brook

- Timber - Timberland
- Veenker - Veenker

A.1.9 *Condition 1 (Nominal)*. : Proximity to various conditions

- Artery - Adjacent to arterial street
- Feedr - Adjacent to feeder street
- Norm - Normal
- RRNn - Within 200 feet of North-South Railroad
- RRAAn - Adjacent to North-South Railroad
- PosN - Near positive off-site feature-park, greenbelt, etc.
- PosA - Adjacent to positive off-site feature
- RRNe - Within 200 feet of East-West Railroad
- RRAe - Adjacent to East-West Railroad

A.1.10 *Condition 2 (Nominal)*. : Proximity to various conditions  
(if more than one is present)

- Artery - Adjacent to arterial street
- Feedr - Adjacent to feeder street
- Norm - Normal
- RRNn - Within 200 feet of North-South Railroad
- RRAAn - Adjacent to North-South Railroad
- PosN - Near positive off-site feature-park, greenbelt, etc.
- PosA - Adjacent to positive off-site feature
- RRNe - Within 200 feet of East-West Railroad
- RRAe - Adjacent to East-West Railroad

A.1.11 *Bldg Type (Nominal)*. : Type of dwelling

- 1Fam - Single-family Detached
- 2FmCon - Two-family Conversion; originally built as one-family dwelling
- Duplx - Duplex
- Twnhse - Townhouse End Unit
- TwnhsI - Townhouse Inside Unit

A.1.12 *Variable: MS Zoning*. MS Zoning (Nominal): Identifies the general zoning classification of the sale.

- A - Agriculture
- C - Commercial
- FV - Floating Village Residential
- I - Industrial
- RH - Residential High Density
- RL - Residential Low Density
- RP - Residential Low Density Park
- RM - Residential Medium Density

A.1.13 *House Style (Nominal)*. : Style of dwelling

- 1Story - One story
- 1.5Fin - One and one-half story: 2nd level finished
- 1.5Unf - One and one-half story: 2nd level unfinished
- 2Story - Two story
- 2.5Fin - Two and one-half story: 2nd level finished
- 2.5Unf - Two and one-half story: 2nd level unfinished
- SFoyer - Split Foyer
- SLvl - Split Level

A.1.14 *Overall Qual (Ordinal)*. : Rates the overall material and finish of the house

- 10 - Very Excellent
- 9 - Excellent

- 8 - Very Good
- 7 - Good
- 6 - Above Average
- 5 - Average
- 4 - Below Average
- 3 - Fair
- 2 - Poor
- 1 - Very Poor

A.1.15 *Overall Cond (Ordinal)*. : Rates the overall condition of the house

- 10 - Very Excellent
- 9 - Excellent
- 8 - Very Good
- 7 - Good
- 6 - Above Average
- 5 - Average
- 4 - Below Average
- 3 - Fair
- 2 - Poor
- 1 - Very Poor

A.1.16 *Roof Style (Nominal)*. : Type of roof

- Flat - Flat
- Gable - Gable
- Gambrel - Gabrel (Barn)
- Hip - Hip
- Mansard - Mansard
- Shed - Shed

A.1.17 *Roof Matl (Nominal)*. : Roof material

- ClyTile - Clay or Tile
- CompShg - Standard (Composite) Shingle
- Membran - Membrane
- Metal - Metal
- Roll - Roll
- Tar and Grv - Gravel and Tar
- WdShake - Wood Shakes
- WdShngl - Wood Shingles

A.1.18 *Exterior 1 and 2 (Nominal)*. : Exterior covering on house

- AsbShng - Asbestos Shingles
- AsphShn - Asphalt Shingles
- BrkComm - Brick Common
- BrkFace - Brick Face
- CBlock - Cinder Block
- CemntBd - Cement Board
- HdBoard - Hard Board
- ImStucc - Imitation Stucco
- MetalSd - Metal Siding
- Other - Other
- Plywood - Plywood
- PreCast - PreCast
- Stone - Stone
- Stucco - Stucco
- VinylSd - Vinyl Siding
- Wd Sdng - Wood Siding
- WdShing - Wood Shingles

A.1.19 *Mas Vnr Type (Nominal)*. : Masonry veneer type

- BrkCmn - Brick Common
- BrkFace - Brick Face
- CBlock - Cinder Block
- None - None
- Stone - Stone

A.1.20 *Bsmt Cond, Exter Qual and Exter Cond (Ordinal)*. : Evaluates the quality of the material on the exterior

- Ex - Excellent
- Gd - Good
- TA - Average/Typical
- Fa - Fair
- Po - Poor

A.1.21 *Foundation (Nominal)*. : Type of foundation

- BrkTil - Brick and Tile
- CBlock - Cinder Block
- PConc - Poured Concrete
- Slab - Slab
- Stone - Stone
- Wood - Wood

A.1.22 *Bsmt Qual (Ordinal)*. : Evaluates the height of the basement

- Ex - Excellent (100+ inches)
- Gd - Good (90-99 inches)
- TA - Typical (80-89 inches)
- Fa - Fair (70-79 inches)
- Po - Poor (<70 inches)
- NA - No Basement

A.1.23 *Bsmt Exposure (Ordinal)*. : Refers to walkout or garden level walls

- Gd - Good Exposure
- Av - Average Exposure (split levels or foyers typically score average or above)
- Mn - Minimum Exposure
- No - No Exposure
- NA - No Basement

A.1.24 *BsmtFin Type 1 (Ordinal)*. : Rating of basement finished area

- GLQ - Good Living Quarters
- ALQ - Average Living Quarters
- BLQ - Below Average Living Quarters
- Rec - Average Rec Room
- LwQ - Low Quality
- Unf - Unfinished
- NA - No Basement

A.1.25 *BsmtFin Type 2 (Ordinal)*. : Rating of basement finished area (if multiple types)

- GLQ - Good Living Quarters
- ALQ - Average Living Quarters
- BLQ - Below Average Living Quarters
- Rec - Average Rec Room
- LwQ - Low Quality

- Unf - Unfinished
- NA - No Basement

A.1.26 *Heating (Nominal)*. : Type of heating

- Floor - Floor Furnace
- GasA - Gas forced warm air furnace
- GasW - Gas hot water or steam heat
- Grav - Gravity furnace
- OthW - Hot water or steam heat other than gas
- Wall - Wall furnace

A.1.27 *Electrical (Ordinal)*. : Electrical system

- SBrkr - Standard Circuit Breakers and Romex
- FuseA - Fuse Box over 60 AMP and all Romex wiring (Average)
- FuseF - 60 AMP Fuse Box and mostly Romex wiring (Fair)
- FuseP - 60 AMP Fuse Box and mostly knob and tube wiring (poor)
- Mix - Mixed

A.1.28 *HeatingQC (Ordinal)*. : Heating quality and condition

- Ex - Excellent
- Gd - Good
- TA - Average/Typical
- Fa - Fair
- Po - Poor

A.1.29 *Central Air (Nominal)*. : Central air conditioning

- N - No
- Y - Yes

A.1.30 *KitchenQual (Ordinal)*. : Kitchen quality

- Ex - Excellent
- Gd - Good
- TA - Typical/Average
- Fa - Fair
- Po - Poor

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and the Teaching Assistants for their support and great suggestions. Authors would also want to thank Kaggle Website and the developers for their valuable information, ideas and contributions.

## REFERENCES

- [1] AiO. 2017. House Prices: Advanced Regression Techniques. (Feb. 2017). <https://www.kaggle.com/notapple/detailed-exploratory-data-analysis-using-r>
- [2] Tanner Carbonati. 2017. Detailed Data Analysis & Ensemble Modeling. (Aug. 2017). <https://www.kaggle.com/tannercarbonati/detailed-data-analysis-ensemble-modeling/notebook>
- [3] Yeshwant Chillakuru, Michael Arango, Jack Crum, and Paul Brewster. 2017. Using Neighborhood Level Data to Predict the Residential Sale Price of Properties in Ames, Iowa. (May 2017). <https://rpubs.com/jackcrum/281471>
- [4] Aarshay Jain. 2016. Complete Guide to Parameter Tuning in XG-Boost. (March 2016). <https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>
- [5] Selva Prabhakaran. 2017. Outlier Treatment. (Dec. 2017). <http://r-statistics.co/Outlier-Treatment-With-R.html>
- [6] Siddharth Raina. 2017. Regularized Regression - Housing Pricing. (Jan. 2017). <https://www.kaggle.com/sidraina89/regularized-regression-housing-pricing>
- [7] Kevin Wong. 2016. Predicting Ames House Prices. (Dec. 2016). <http://kevinfw.com/post/predicting-ames-house-prices/>

- [8] Ricky Yue and Jurgen De Jager. 2016. Advanced Regression Modeling on House Prices. (Sept. 2016). <https://nycdatascience.com/blog/student-works/advanced-regression-modeling-house-prices/>

## LIST OF FIGURES

1	Code - Null Checks	11
2	Graph - Missing Values	11
3	Code - Numerical Analysis	12
4	Graph - Sale Price and Overall Quality	12
5	Graph - Ground Live Area and Year Built	13
6	Code - Categorical Analysis	13
7	Graph - Neighborhood and Sale Type	14
8	Code - Correlations	14
9	Graph - Correlations with Sale Price	15
10	Graph - Overall Quality and Ground Live Area	15
11	Graph - Garage Cars and Garage Area	16
12	Graph - Sale Price skewness	16
13	Code - Outlier Analysis	16
14	Graph - Outliers using Cook's distance	17
15	Graph - Garage Live Area Outliers	17
16	Graph - Garage Area Outliers	18
17	Code - Delete Outliers	18
18	Code - factorize and one-hot encoding	18
19	Code - SVM Algorithm	19
20	Code - Random Forest Algorithm	19
21	Graph - Random Forest Feature Ranking	20
22	Code - Lasso Algorithm	21
23	Code - Ridge Algorithm	22
24	Graph - Ridge Top 10 Positive Features	23
25	Graph - Ridge Top 10 Negative Features	24
26	Code - XGBoost Algorithm	25
27	Graph - XGBoost Feature Importance	26

```

# python code - check for null values
train = pd.read_csv('../data/train.csv')
test = pd.read_csv('../data/test.csv')

#combine the data sets
alldata = train.append(test)
na = alldata.isnull().sum()
    .sort_values(ascending=False)

```

Figure 1: Code - Null Checks

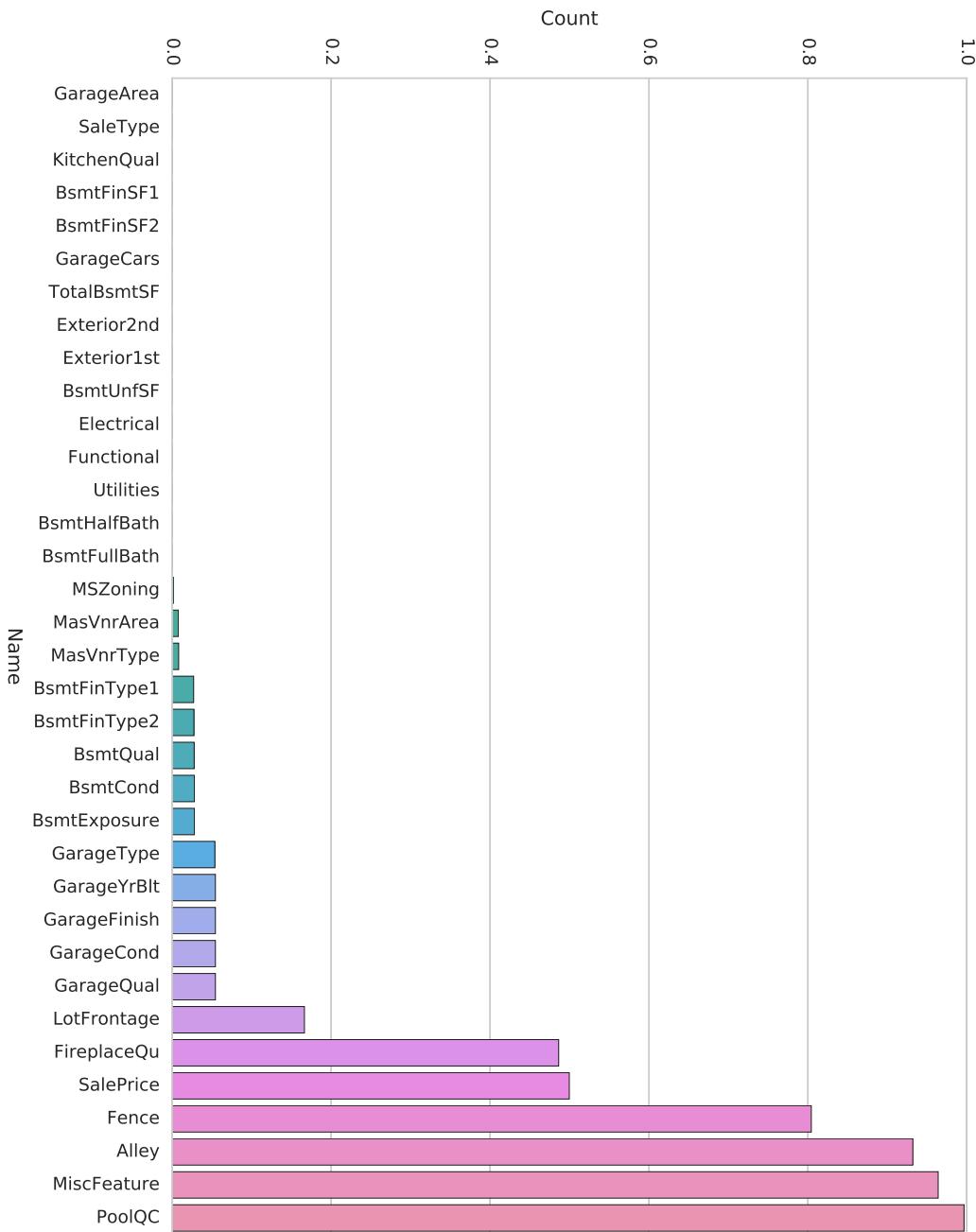


Figure 2: Graph - Missing Values

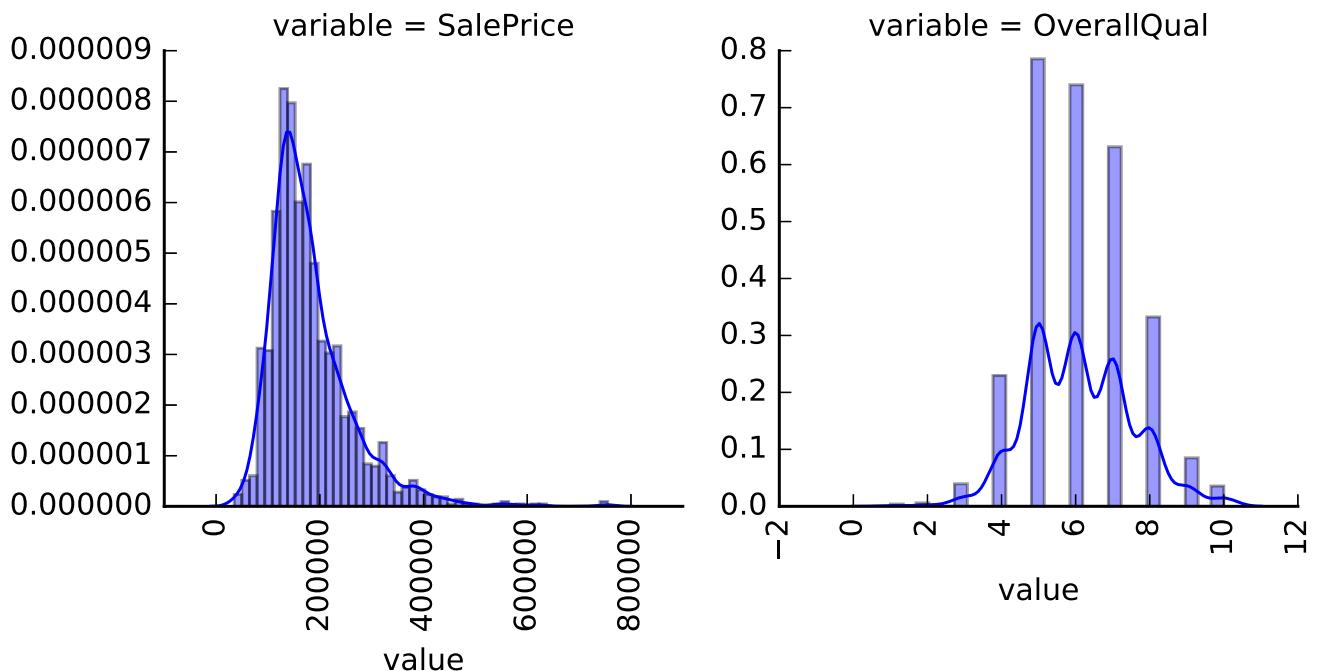
```

# python code - analyze numeric variables
numerical_features = [f for f in train.columns
if train.dtypes[f] != 'object']

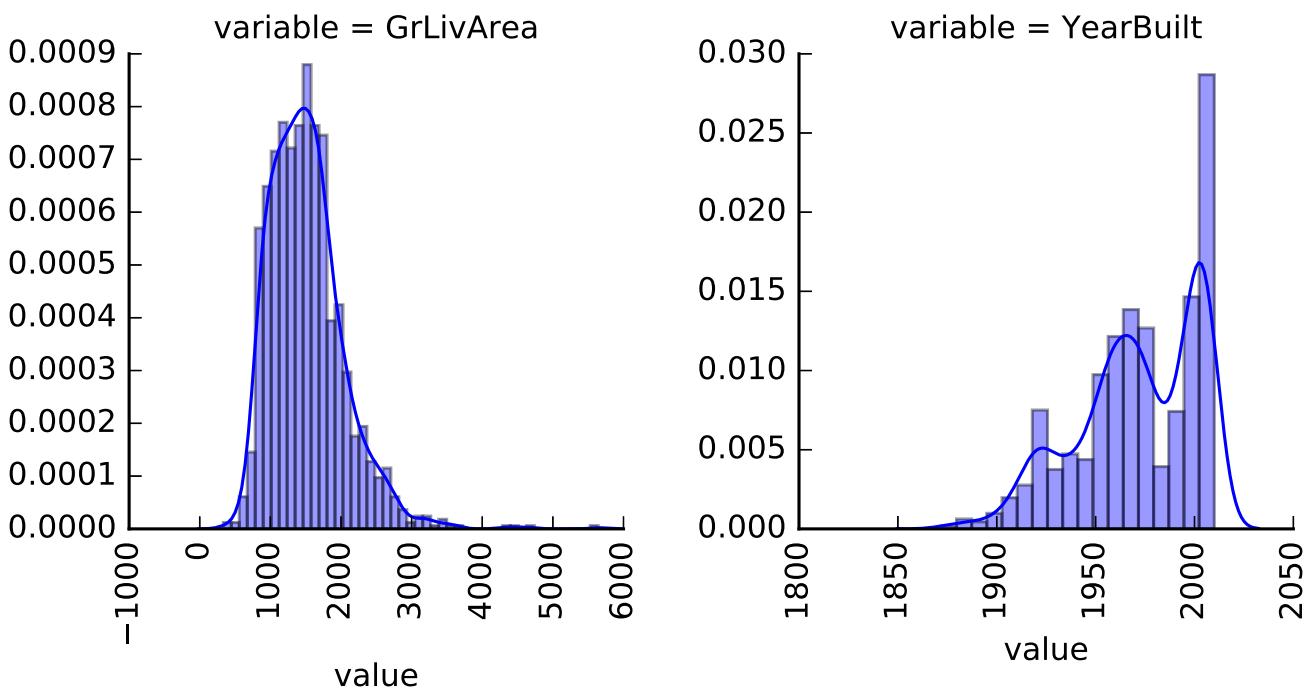
nd = pd.melt(train, value_vars = numerical_features)
plt.figure(figsize = (5,3))
plot = sns.FacetGrid (nd, col='variable', col_wrap=4,
                     sharex=False, sharey = False)
plot = plot.map(sns.distplot, 'value')

```

**Figure 3: Code - Numerical Analysis**



**Figure 4: Graph - Sale Price and Overall Quality**



**Figure 5: Graph - Ground Live Area and Year Built**

```
# python code - analyze numeric variables
cat_features = [f for f in train.columns
if train.dtypes[f] == 'object']
print(cat_features)

def barplot(x,y,**kwargs):
    sns.barplot(x=x,y=y)
    x = plt.xticks(rotation=90)

plt.figure(figsize = (5,3))

p = pd.melt(train, id_vars='SalePrice',
            value_vars=cat_features)

g = sns.FacetGrid (p, col='variable', col_wrap=4,
sharex=False, sharey=False, size=5)

g = g.map(barplot, 'value','SalePrice')
```

**Figure 6: Code - Categorical Analysis**

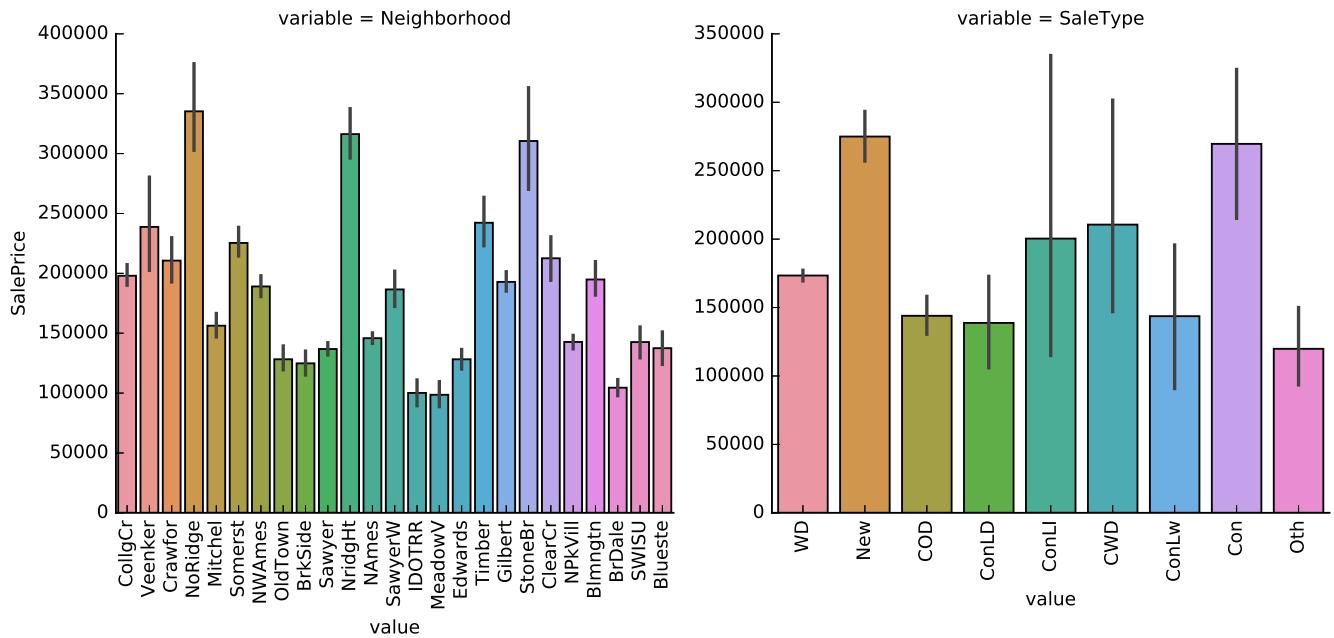


Figure 7: Graph - Neighborhood and Sale Type

```
# python code
corr = alldata[numerical_features].corr()
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
plt.figure(figsize = (15,8))
sns_plot = sns.heatmap(corr, cmap='YlGnBu',
                      linewidths=.5, mask=mask, vmax=.3)
```

Figure 8: Code - Correlations

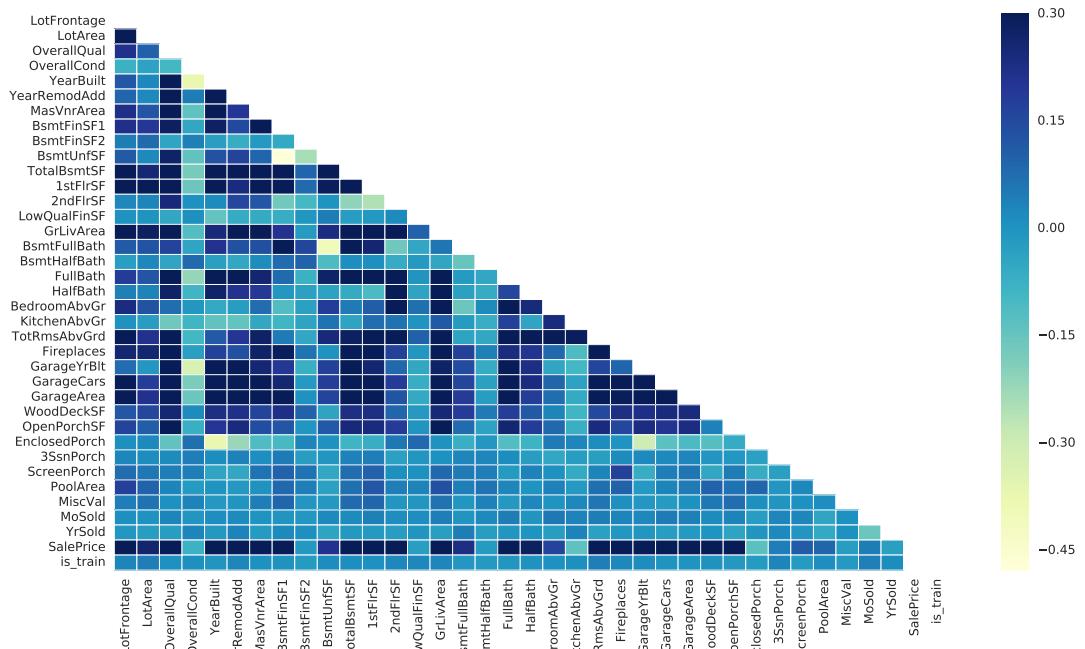


Figure 9: Graph - Correlations with Sale Price

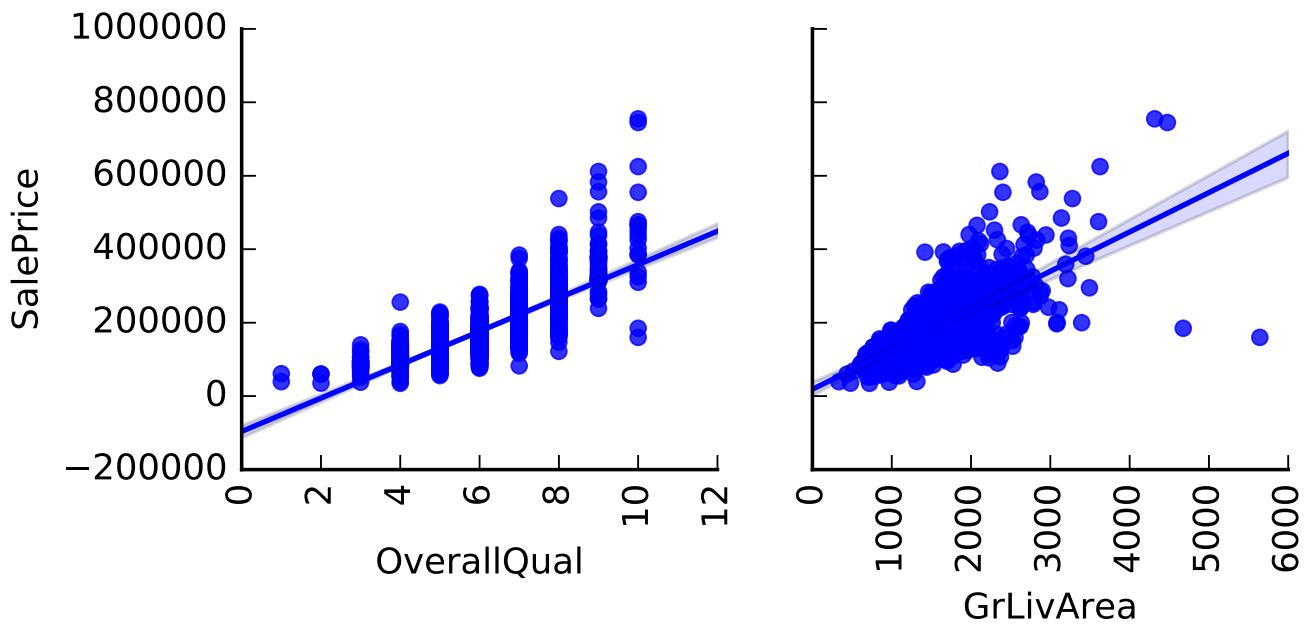


Figure 10: Graph - Overall Quality and Ground Live Area

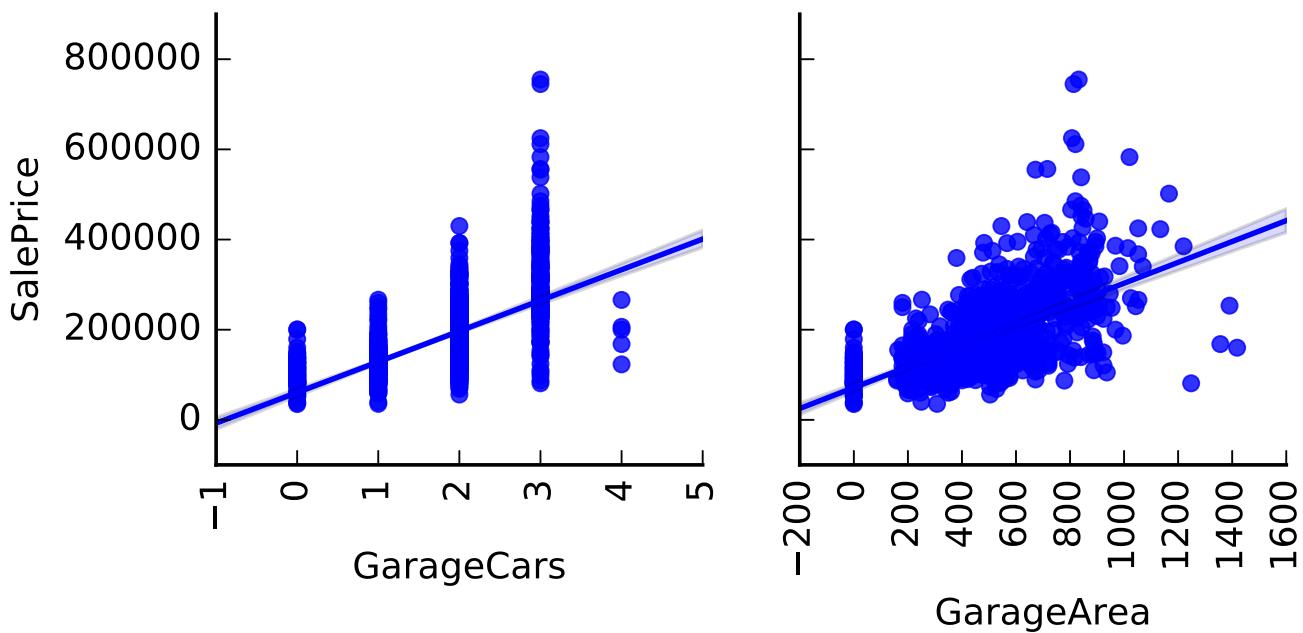


Figure 11: Graph - Garage Cars and Garage Area

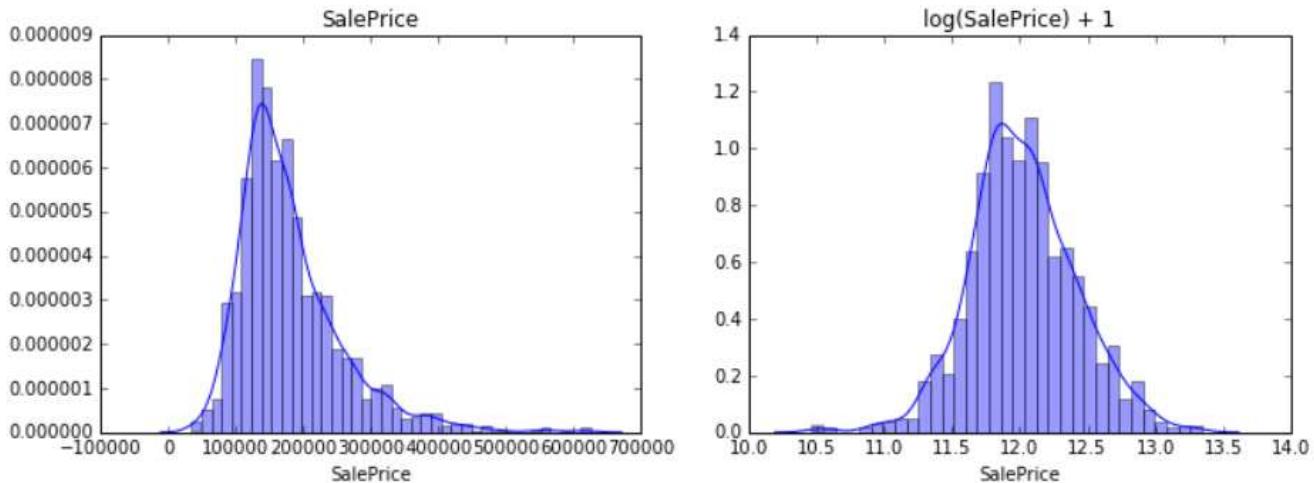
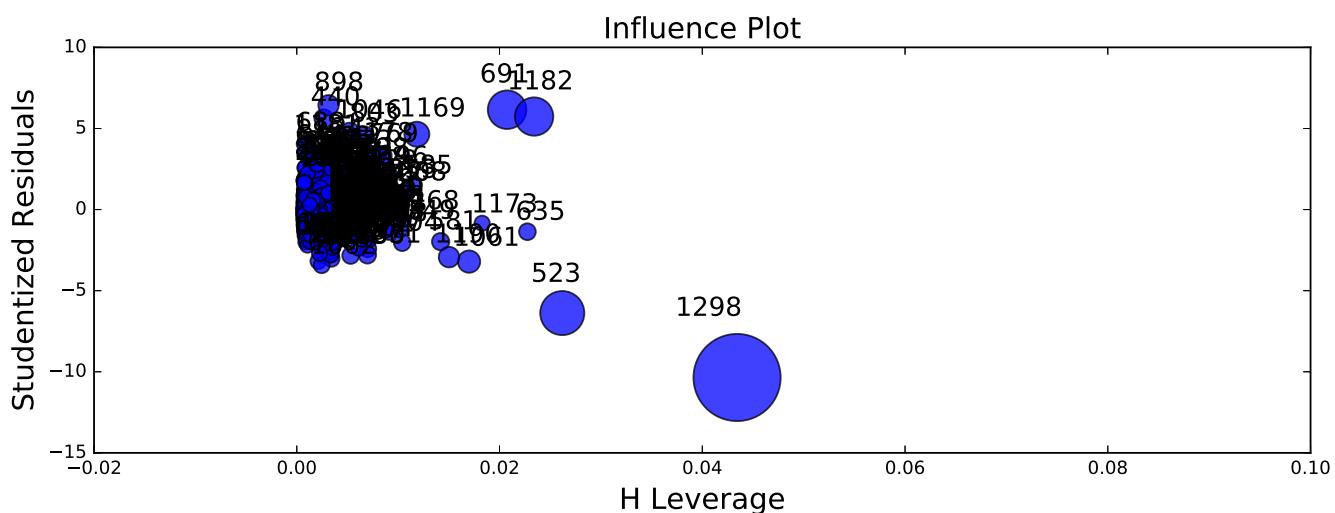


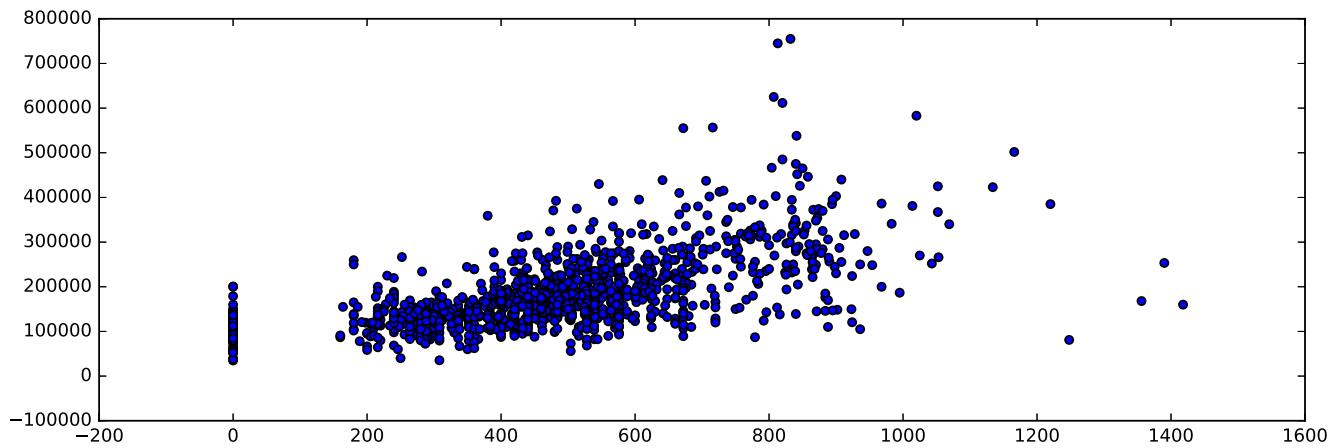
Figure 12: Graph - Sale Price skewness

```
# python code - outlier analysis
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols(formula = 'SalePrice ~
GrLivArea + GarageArea', data=train)
fitted = model.fit()
plot = sm.graphics.influence_plot(fitted,
criterion='cooks')
```

Figure 13: Code - Outlier Analysis





**Figure 16: Graph - Garage Area Outliers**

```
# python code - remove outlier rows
# fix all extreme outliers based on outlier analysis
# 8 rows will be deleted
train = train[train.GrLivArea <= 4000]
train = train[train.GarageArea <= 1200]
```

**Figure 17: Code - Delete Outliers**

```
# python code - factorize and one-hot
def get_one_hot(df, col_name, fill_val):
    if fill_val is not None:
        df[col_name].fillna(fill_val, inplace=True)

    dummies = pd.get_dummies(df[col_name], prefix='_' + col_name)
    df = df.join(dummies)
    df = df.drop([col_name], axis=1)
    return df
#end def

from sklearn.preprocessing import LabelEncoder

def factorize(df, column, fill_na=None):
    le = LabelEncoder()
    if fill_na is not None:
        df[column].fillna(fill_na, inplace=True)
    le.fit(df[column].unique())
    df[column] = le.transform(df[column])
    return df
#end def
```

**Figure 18: Code - factorize and one-hot encoding**

```

# python code - SVM algorithm
from sklearn.svm import SVR
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error

train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
target_vector = train['SalePrice']

_svm_algo = SVR(kernel = 'rbf', C=1e3, gamma=1e-8)

_svm_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _svm_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _svm_algo.predict(test)

```

**Figure 19: Code - SVM Algorithm**

```

# python code - random forest algorithm
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import mean_squared_error

train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
target_vector = train['SalePrice']

_algo = RandomForestRegressor(n_estimators=100,
oob_score=True, random_state=123456)

model = _algo.fit(train, target_vector)

feat_imp = pd.Series(_algo.feature_importances_,
train.columns).sort_values(ascending=False)

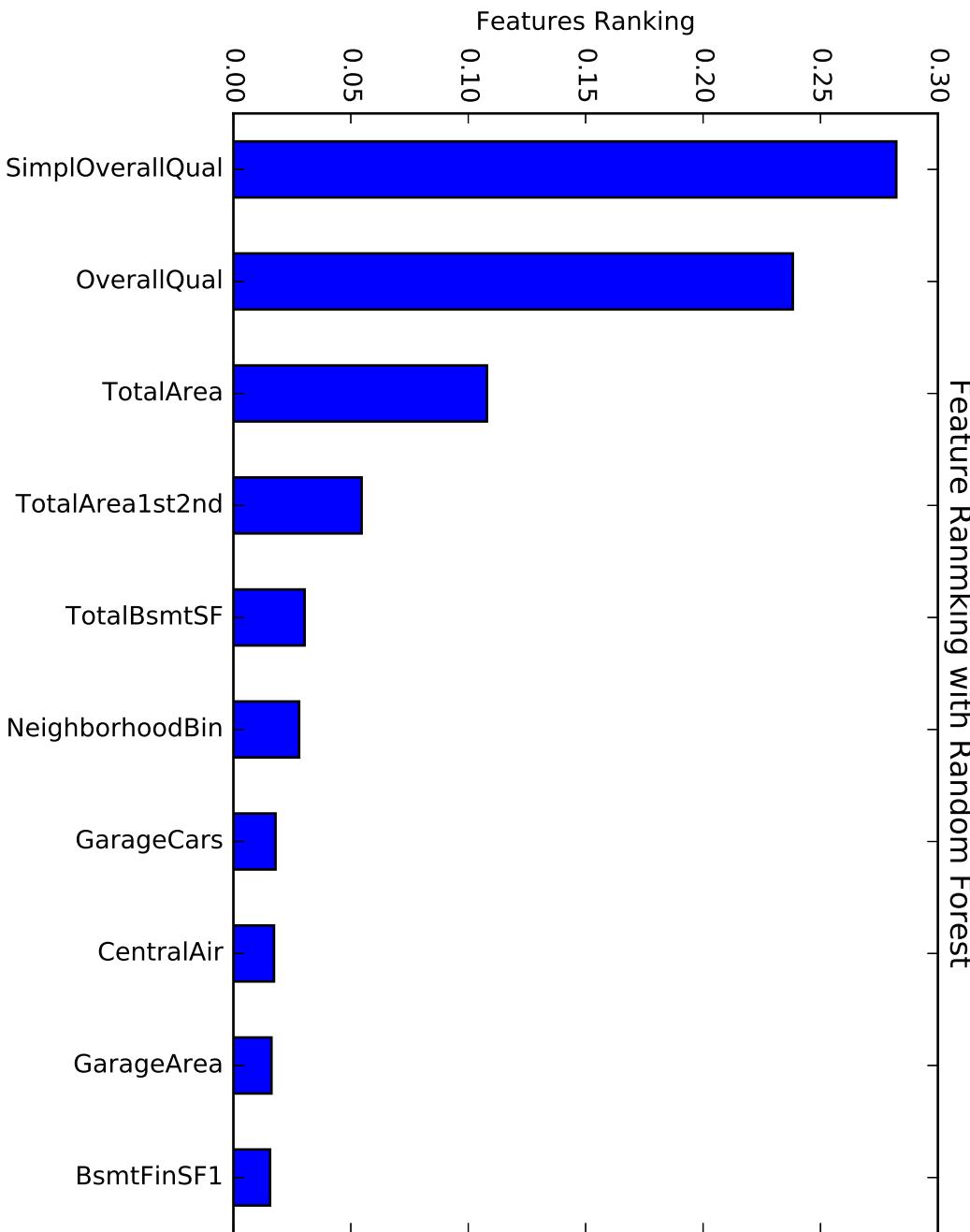
feat_imp[:10].plot(kind='bar',
title='Feature Ranmkingt')
y_train = target_vector
y_train_pred = _algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))
y_test_pred = _algo.predict(test)

```

**Figure 20: Code - Random Forest Algorithm**

Figure 21: Graph - Random Forest Feature Ranking



```

# python code - lasso algorithm
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_squared_error

train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
target_vector = train['SalePrice']

#found this best alpha value through cross-validation
_best_alpha = 0.0001

_lasso_algo = Lasso(alpha = _best_alpha,
                     max_iter = 50000)

model = _lasso_algo.fit(train, target_vector)

y_train = target_vector
y_train_pred = _algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _lasso_algo.predict(test)

```

**Figure 22: Code - Lasso Algorithm**

```

# python code - ridge algorithm
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error

train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
target_vector = train['SalePrice']

#found this best alpha value through cross-validation
_best_alpha = 0.00099

_ridge_algo = Ridge(alpha = _best_alpha,
                     normalize = True)

_ridge_algo.fit(train, target_vector)

df = {'features': train.columns.values,
       'Coefficients': _ridge_algo.coef_[0]}
coefficients = pd.DataFrame(df)
           .sort_values(by='Coefficients',
                        ascending=False)

plt.figure()
coefficients.iloc[0:10].plot(x=['features'],
                             kind='bar', title='Top 10 Positive Features')
plt.ylabel('Feature Coefs')
plt.figure()
coefficients.iloc[-10: ].plot(x=['features'],
                               kind='bar', title='Top 10 Negative Features')
plt.ylabel('Feature Coefs')

y_train = target_vector
y_train_pred = _ridge_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _ridge_algo.predict(test)

df_predict = pd.DataFrame({'Id': test['Id'],
                           'SalePrice': np.exp(y_test_pred) - 1.0})

#df_predict = pd.DataFrame({'Id': id_vector,
#                           'SalePrice': sale_price_vector})

df_predict.to_csv('../data/kaggle_python_ridge.csv',
                 header=True, index=False)

```

**Figure 23: Code - Ridge Algorithm**

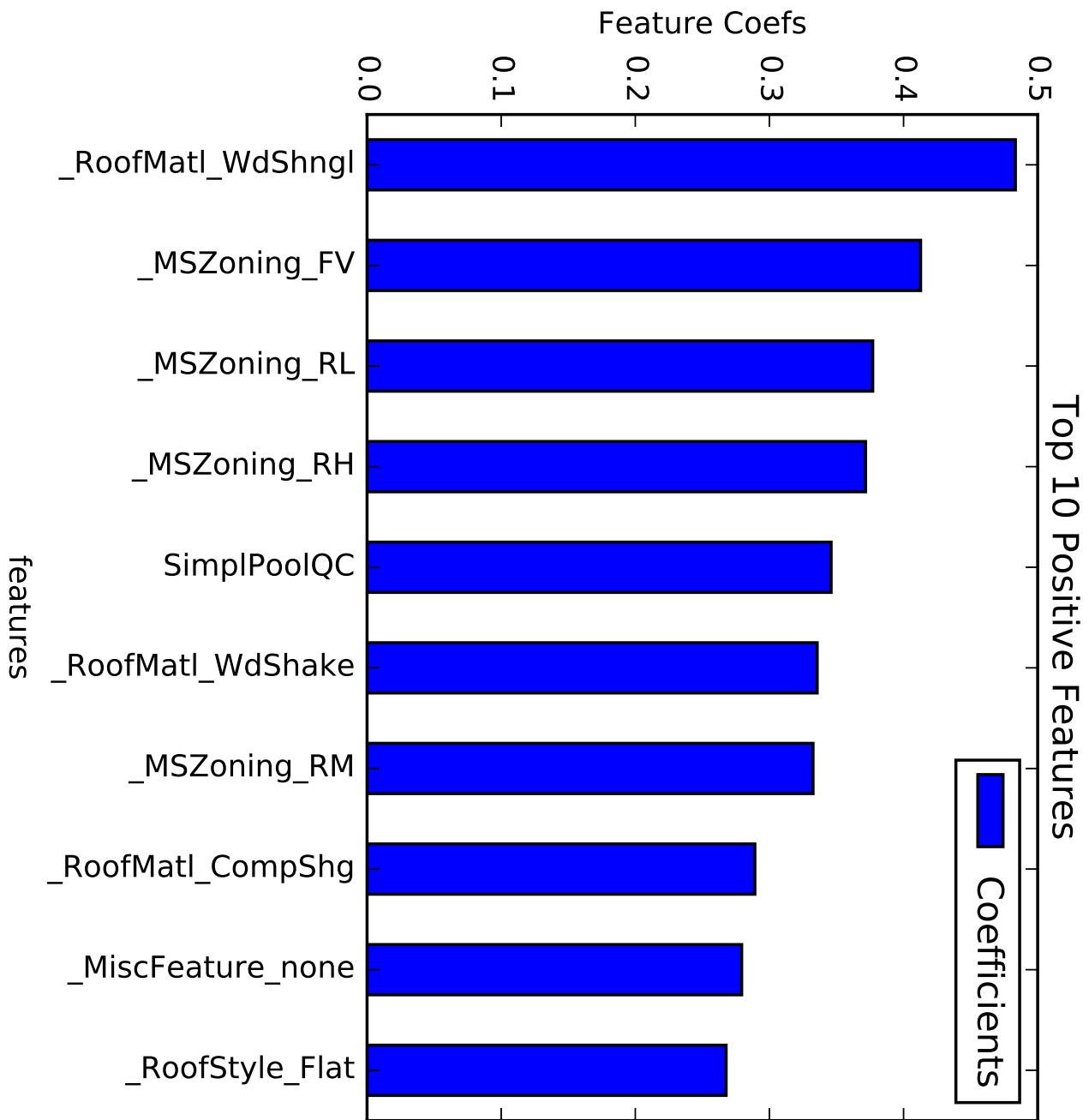


Figure 24: Graph - Ridge Top 10 Positive Features

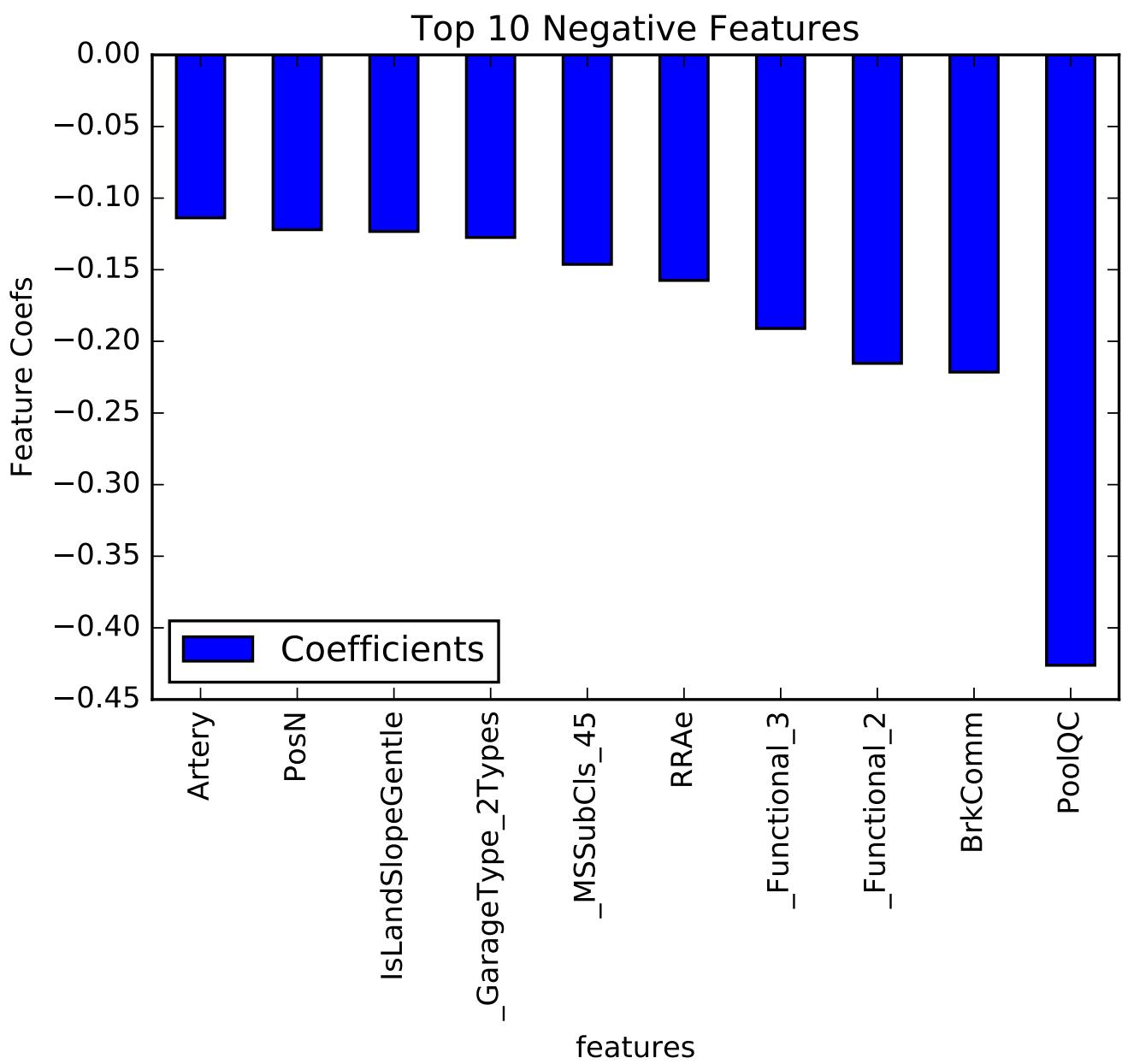


Figure 25: Graph - Ridge Top 10 Negative Features

```

# python code - XGBoost algorithm
import xgboost as xgb
from xgboost import XGBClassifier
from xgboost import plot_importance
from sklearn.metrics import mean_squared_error
from sklearn import cross_validation, metrics

train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
target_vector = train['SalePrice']

_algo = np.random.seed(1234)

_xgb_algo = xgb.XGBRegressor(
    colsample_bytree=0.8,
    colsample_bylevel = 0.8,
    gamma=0.01,
    learning_rate=0.05,
    max_depth=5,
    min_child_weight=1.5,
    n_estimators=6000,
    reg_alpha=0.5,
    reg_lambda=0.5,
    subsample=0.7,
    seed=42,
    silent=1)

_xgb_algo.fit(train, target_vector)

feat_imp = pd.Series(_xgb_algo.booster()
    .get_fscore())
.sort_values(ascending=False)[0:10]
plot = feat_imp.plot(kind='bar',
    title='Top 10 Feature Importances')

y_train = target_vector
y_train_pred = _xgb_algo.predict(train)

#root mean squared error (RMSE)
rmse_train = np.sqrt(rmse(y_train,y_train_pred))

y_test_pred = _xgb_algo.predict(test)

```

**Figure 26: Code - XGBoost Algorithm**

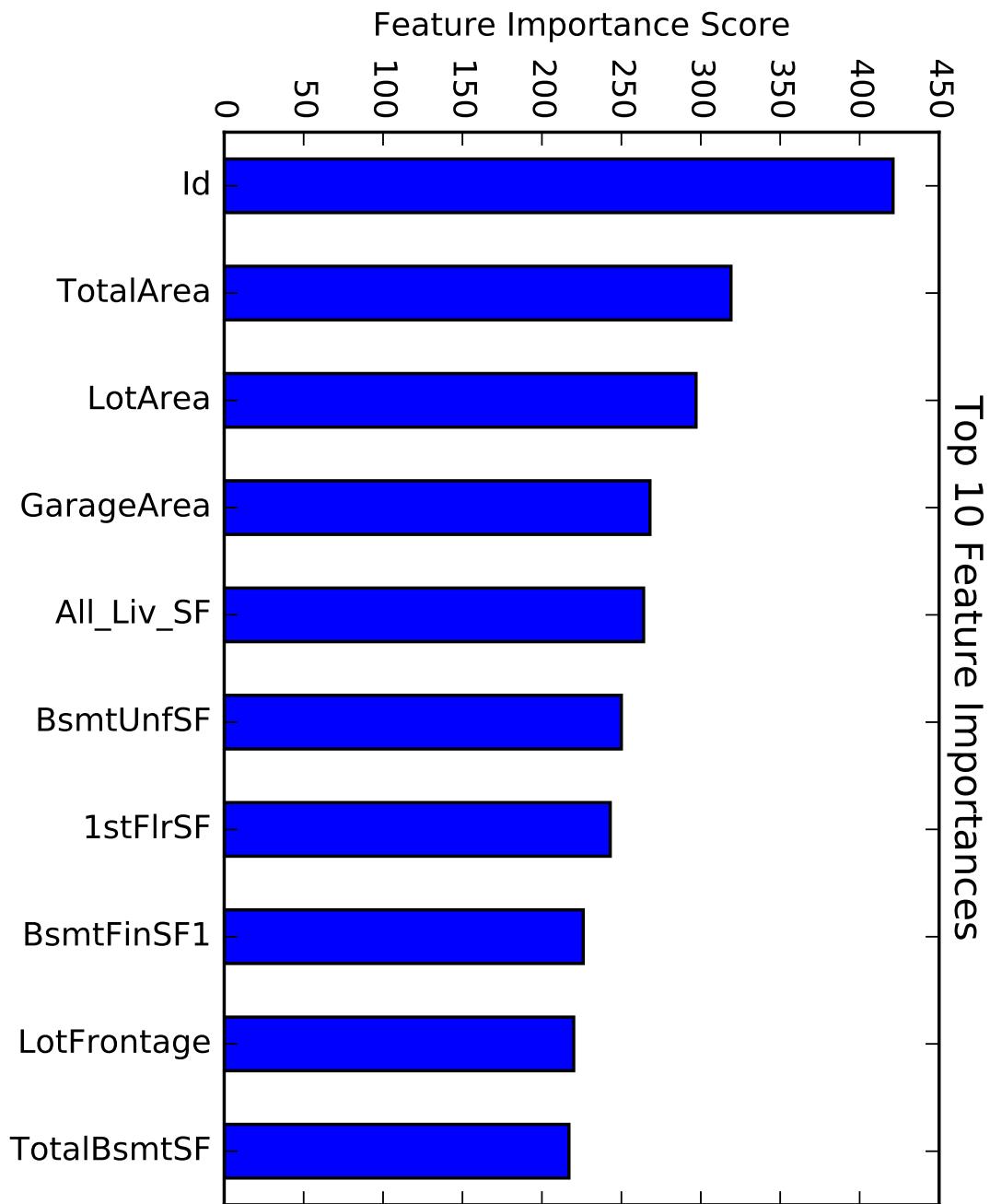


Figure 27: Graph - XGBoost Feature Importance

LIST OF TABLES

1	Table - XGBoost Hyper-parameters	28
2	Table - Kaggle Submissions	28

**Table 1: Table - XGBoost Hyper-parameters**

Hyper-parameter	Description
Maximum Iterations:	Number of trees in the final model. More the trees, more accuracy.
Maximum Depth:	Depth of each individual tree to control overfitting.
Step Size:	Shrinkage, works similar to learning rate; smaller value takes more iterations.
Column Subsample:	Subset of the columns to use in each iteration.

**Table 2: Table - Kaggle Submissions**

Algorithm	RMSE	Kaggle Score
SVM	0.2069	0.23967
Random Forest	0.0519	0.14607
Ridge	0.0988	0.13687
XGBoost	0.0432	0.13018
Lasso	0.1015	0.12860
Neural Network	0.20	0.12510
Ensemble		0.12011

```
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
```

---

```
[2017-12-10 13.50.33] 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 1.2s.
```

---

```
Compliance Report
```

---

```
name: Cheruvu, Murali
hid: 306
paper1: 100%; 10/26/2017
paper2: 100%; 11/4/2017
project: 100%; 12/3/2017
```

```
yamlcheck
```

---

```
wordcount
```

---

```
28
```

```
wc 306 project 28 7041 report.tex  
wc 306 project 28 7419 report.pdf  
wc 306 project 28 203 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

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

---

```
passed: False
```

```
floats
```

---

```
66: First part of the analysis was to check for any missing values in  
the training and testing datasets as shown in figure  
(\ref{c:check-nulls}). Using the bar plot shown in figure  
(\ref{fig:missing-values}), we have identified that there are 5  
variables: {\em pool quality}, {\em miscellaneous features}, {\em  
alley}, {\em fence} and {\em fire place quality}, having the most  
missing data. All the missed values for the numeric variables are  
analyzed further to decide whether we need to delete the instances  
of all the data with missing values or impute them using  
meaningful data such as median of the corresponding variable.
```

```
68: \begin{figure}[htb]
```

```
79: \caption{Code - Null Checks}\label{c:check-nulls}
```

```
82: \begin{figure}[htb]
```

```

84: \includegraphics[width=1.0\columnwidth]{images/missing_values}
85: \caption{Graph - Missing Values} \label{fig:missing-values}
89: There are 37 numerical variables after excluding the {\em Id} variable. We have analyzed all the numerical variables for data patterns such as skewness in the data and range of the possible values. We have shown {\em sale price}, {\em overall quality}, {\em garage live area} and {\em year built}, in the figures (\ref{fig:num-feature-1}) and (\ref{fig:num_features_2}) as a few sample plots from the numerical analysis. Corresponding code snippet is shown in figure (\ref{c:analyze-numeric}).
91: \begin{figure}[htb]
103: \caption{Code - Numerical Analysis}\label{c:analyze-numeric}
107: \begin{figure}[htb]
109: \includegraphics[width=1.0\columnwidth]{images/num_features_1}
110: \caption{Graph - Sale Price and Overall Quality}\label{fig:num-feature-1}
113: \begin{figure}[htb]
115: \includegraphics[width=1.0\columnwidth]{images/num_features_2}
116: \caption{Graph - Ground Live Area and Year Built}\label{fig:num_features_2}
120: There are 43 categorical variables in the combined dataset. We have analyzed all categorical variables and found the ways to fill the missing values. We has also evaluated proper approaches to convert them into numerical factors. Later on in the feature engineering section, we will go through more details on numerical factors. Categorical variable factors and the corresponding code snippet for {\em neighborhood} and {\em sale type} are shown in figure (\ref{c:analyze-cat}) and figures (\ref{fig:cat_features_1}).
122: \begin{figure}[htb]
143: \caption{Code - Categorical Analysis} \label{c:analyze-cat}
146: \begin{figure}[htb]
148: \includegraphics[width=1.0\columnwidth]{images/cat_features_1}
149: \caption{Graph - Neighborhood and Sale Type}\label{fig:cat_features_1}
153: {\em Numpy} package offers correlations functionality to analyze the variables that are positively or negatively correlated with the {\em sale price} and also analyze any interdependencies among the variables. Figure (\ref{c:cor}) and (\ref{fig:correlations}) shows the code snippet and the correlations plot. From that we can list the top 10 features those are strongly correlated with the target variable - {\em sale price}. We can visualize a few pair-wise correlation graphs with sale price for further detailed analysis. Figures (\ref{fig:pair-wise-correlations}) and (\ref{fig:pair-wise-correlations-2}) show how {\em overall quality}, {\em ground live area}, {\em garage cars} and {\em

```

garage area} are positively correlated with {\em sale price}.  
 155: \begin{figure}[htb]  
 165: \caption{Code - Correlations} \label{c:cor}  
 168: \begin{figure}[htb]  
 170: \includegraphics[width=1.0\columnwidth]{images/correlations}  
 171: \caption{Graph - Correlations with Sale Price}  
     \label{fig:correlations}  
 174: \begin{figure}[htb]  
 176: \includegraphics[width=1.0\columnwidth]{images/pair\_wise\_correlat  
     ions\_1}  
 177: \caption{Graph - Overall Quality and Ground Live Area}  
     \label{fig:pair-wise-correlations}  
 180: \begin{figure}[htb]  
 182: \includegraphics[width=1.0\columnwidth]{images/pair\_wise\_correlat  
     ions\_2}  
 183: \caption{Graph - Garage Cars and Garage Area} \label{fig:pair-  
     wise-correlations-2}  
 200: From the numerical analysis, we have identified that there are a  
     few numerical variables need further analysis to identify the  
     skewed data. We did not find any key variables those have skewed  
     more than 75\%. However, we wanted to replace the {\em sale  
     price} with corresponding logarithmic value for the predictive  
     models and later convert it back to the exponential value before  
     submitting to the kaggle competition. Figure (\ref{fig:sale-  
     price-skew}) shows the {\em sale price}, before and after  
     applying the logarithmic value.  
 202: \begin{figure}[htb]  
 204: \includegraphics[width=1.0\columnwidth]{images/sale\_price\_skew}  
 205: \caption{Graph - Sale Price skewness} \label{fig:sale-price-skew}  
 210: Continuing with exploratory analysis, we have analyzed the  
     outliers using {\em Cook's distance}. Cook's distance is a  
     measure calculated from a regression model to find out the  
     influence exerted by each observation (row) on the predictions.  
     As a practice, those observations that have a Cook's distance  
     greater than 4 times the mean value may be classified as an  
     outlier. Outlier detection can be done using univariate and  
     multivariate analysis. In univariate model, the outliers are  
     those observations that are present outside of  $1.5 * \text{IQR}$ , where  
     IQR ({\em Inter Quartile Range}) is the difference between 75th  
     and 25th quartiles. Analyzing outliers in any observations based  
     on single variable may lead to incorrect inferences. Cook's  
     distance generalizes the outlier analysis using multivariate  
     approach \cite{1}. Figure (\ref{c:code-outliers}) is the code  
     implementing Cook's distance to find the outliers from training  
     dataset and figure (\ref{fig:outliers}) shows the scatter plot  
     with outliers being marked as bubbles. The bigger the bubbles,

the bigger outlier deviations from the mean value. We have further analyzed two key variables - {\em ground live area} and {\em garage area} that are in high correlation with the {\em sale price}. From the scatter plot shown in figure (\ref{fig:gr-liv-area-outlier}), we can see that {\em garage live area} has 4 outliers with values greater than 4,000 sq ft. We can also visualize 4 outliers in {\em garage area} scatter plot with values greater than 1,200 sq ft. as shown in figure (\ref{fig:garage-area-outlier}). We have removed the 8 outlier rows related to these two variables from the training dataset, the corresponding code snippet shown in figure (\ref{c:code-del-outliers}).

```

212: \begin{figure}[htb]
224: \caption{Code - Outlier Analysis} \label{c:code-outliers}
227: \begin{figure}[htb]
229: \includegraphics[width=1.0\columnwidth]{images/outliers}
230: \caption{Graph - Outliers using Cook's distance}
    \label{fig:outliers}
233: \begin{figure}[htb]
235: \includegraphics[width=1.0\columnwidth]{images/gr_liv_area_outlie
r}
236: \caption{Graph - Garage Live Area Outliers} \label{fig:gr-liv-
area-outlier}
239: \begin{figure}[htb]
241: \includegraphics[width=1.0\columnwidth]{images/garage_area_outlie
r}
242: \caption{Graph - Garage Area Outliers} \label{fig:garage-area-
outlier}
245: \begin{figure}[htb]
253: \caption{Code - Delete Outliers} \label{c:code-del-outliers}
273: One-hot encoding converts the category variable into many binary vectors, one new numeric variable for each value in the category. Assume that we have a categorical variable called signal-light with three possible values: green, yellow and red. We will need to convert these values into numeric - green = 1, yellow = 2 and red = 3. When we apply one-hot encoding on this variable, basically we are creating three new categorical variables - signal-light-green, signal-light-yellow and signal-light-red along with the original variable - signal-light, each is pretty much a binary vector having 1s for all the corresponding values; otherwise 0s. With hot-encoding, we are basically increasing dimensions in the model. After extensive feature engineering applied on the housing dataset, we have added {\em 228} new features (variables). Figure (\ref{c:code-one-hot}) shows the python methods to factorize categorical variables and one-hot encoding techniques.

```

```

275: \begin{figure}[htb]
299: \caption{Code - factorize and one-hot encoding} \label{c:code-one-hot}
323: Support Vector Machine (SVM) algorithms can be used to solve
      classification and regression problems. SVM regression relies on
      kernel functions for modeling the data. SVM creates larger
      margins between categories of data so that they are linearly
      separable. SVM handles non-linearly separable data, mainly for
      regression problems, using kernel functions, such as polynomial,
      radial basis function (RBF) and sigmoid, to project the data onto
      a hyperplane. Figure (\ref{c:svm}) shows the python
      implementation for {\em sale price} predictions of the housing
      test dataset.
325: \begin{figure}[htb]
349: \caption{Code - SVM Algorithm} \label{c:svm}
355: Random Forest is an advanced machine learning algorithm for
      predictive analytics. Random Forest ensembles multiple decision
      trees to create an additive learning model from the sequence of
      base models created by each decision tree that worked on a sub-
      sample dataset. Random Forest models are suitable to handle
      tabular datasets with hundreds of numeric and categorical
      features. Along with missing values, non-linear relations between
      features and the target, will be handled well by random forest
      algorithms. With proper tuning of hyper-parameters of the random
      forest algorithm, it can perform well with decent accuracy in the
      predictions without overfitting the model. Unlike similar
      regression models, it does not offer feature coefficient
      information but it provides {\em feature ranking} functionality
      very nicely. Figure (\ref{c:rf}) shows the random forest
      algorithm details for the {\em sale price} predictions
      implemented using {\em sklearn} package and the figure
      (\ref{fig:random-feature-ranking}) shows the top 10 important
      features selected by random forest to model the predictions.
357: \begin{figure}[htb]
386: \caption{Code - Random Forest Algorithm} \label{c:rf}
390: \begin{figure}[htb]
392: \includegraphics[width=1.0\columnwidth]{images/random_forest_feature_ranking}
393: \caption{Graph - Random Forest Feature Ranking}
      \label{fig:random-feature-ranking}
398: Lasso is a regression model that uses shrinkage to bring data
      points towards the center, similar to the mean value of all the
      data points. Lasso stands for Least Absolute Shrinkage and
      Selection Operator. It is a regularized linear model with penalty
      term {\em lambda} to minimize the error. Parameter penalization
      controls overfitting the input data by shrinking variable

```

coefficients to 0. Essentially this makes the variables no effect in the model, hence reduces the dimensions. Figure (\ref{c:lasso}) shows the lasso algorithm implementation for \em sale price predictions in python.

```
400: \begin{figure}[htb]
426: \caption{Code - Lasso Algorithm} \label{c:lasso}
431: Ridge algorithm is very similar to lasso algorithm with the same goal. While lasso performs \em L1 regularization, ridge applies \em L2 regularization techniques in modeling the predictions. L1 regularization adds penalty to the variables equivalent to \em absolute value of the magnitude of the coefficients, whereas L2 adds the penalty equivalent to \em square of the magnitude of the variable coefficients. Figure (\ref{c:ridge}) shows the python implementation of the ridge algorithm for the \em sale price predictions. Figures (\ref{fig:ridge-feature-ranking-pos}) and (\ref{fig:ridge-feature-ranking-neg}) show the top 10 positively and top 10 negatively influencing variables with \em sale price.
433: \begin{figure}[htb]
433: \caption{Code - Ridge Algorithm} \label{c:ridge}
486: \begin{figure}[htb]
488: \includegraphics[width=1.0\columnwidth]{images/ridge_feature_ranking_pos}
489: \caption{Graph - Ridge Top 10 Positive Features}
        \label{fig:ridge-feature-ranking-pos}
492: \begin{figure}[htb]
494: \includegraphics[width=1.0\columnwidth]{images/ridge_feature_ranking_neg}
495: \caption{Graph - Ridge Top 10 Negative Features}
        \label{fig:ridge-feature-ranking-neg}
500: XGBoost (eXtreme Gradient Boosting) is one of the Gradient Boosted Machine algorithms. It ensembles (combines) optimized model by taking trained models from all the preceding iterations. XGBoost regularizes the variables (parameters) to reduce the overfit and can work well with variables having missing values. It is empowered with built-in cross validation to reduce the boosting iterations; hence offers better performance along with parallel processing on distributed systems such as Hadoop. By tuning the XGBoost hyper parameters, we can achieve well optimized model that can make more accurate predictions. XGBoost uses \em F-Score to measure the importance of variables. Table (\ref{tab:xgb-param}) explains the hyper-parameters of XGBoost algorithm and also given the python code, as shown in figure (\ref{c:xgb}), implementing for \em sale price predictions. Figure (\ref{fig:xgb-feature-imp}) shows the top 10 feature selection by the XGBoost.
```

```

502: \begin{table}[htb]
505: \label{tab:xgb-param}
521: \begin{figure}[htb]
567: \caption{Code - XGBoost Algorithm} \label{c:xgb}
570: \begin{figure}[htb]
572: \includegraphics[width=1.0\columnwidth]{images/xgboost_feature_im-
portance}
573: \caption{Graph - XGBoost Feature Importance} \label{fig:xgb-
feature-imp}
581: We can create a robust predictive model with better accuracy by
merging two or more machine learning algorithms. This technique
is called {\em model ensembling}. Ensembled algorithms may be
similar in functionality or may entirely be different from each
other. Individual algorithms may not perform great but by
ensembling them, the overall system can offer much better
performance and accuracy. Variations in the predicting logic in
each of these individual algorithms will bring unbiasedness into
the unified model. {\em Bagging}, {\em boosting} and {\em
stacking} are popular ensembling techniques. Many of the advanced
machine learning algorithms use ensembled approaches to achieve
accurate classifications or predictions. Random Forest uses
bagging, XGBoost uses boosting and Neural Network applies
stacking ensembling techniques. For the kaggle submission, we
have created an ensembled model by averaging {\em Sale Price} of
the top 3 performing ensembled algorithms - XGBoost, Lasso and
Neural Network. As predicted, ensembled model has scored better
compared to the individual algorithms. By applying advanced
machine learning algorithms, we have placed our scores within top
20\% of the competition. Table (\ref{tab:kaggle}) displays each
algorithm and the {\em root mean squared error} (RMSE) along with
the {\em kaggle score}.
583: \begin{table}[htb]
585: \label{tab:kaggle}
615: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/mas-
ter/project/code/1.1_exploratory_analysis_numerical.ipynb}{1.1\_e-
xploratory\_analysis\_numerical.ipynb} - To load datasets and
analyze all numerical variables
617: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/mas-
ter/project/code/1.2_exploratory_analysis_categorical.ipynb}{1.2\_e-
xploratory\_analysis\_categorical.ipynb} - To analyze
categorical variables in the dataset
619: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/mas-
ter/project/code/1.3_outlier_and_skewed_data_analysis.ipynb}{1.3\_o-
utlier\_and\_skewed\_data\_analysis.ipynb} - Handles outlier
and skewed data analysis
621: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/mas-

```

ter/project/code/1.4\_feature\_engineering.ipynb}{1.4\\_feature\\_engineering.ipynb} - All the feature engineering is done in this file

623: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.1\_algorithm\_svm.ipynb}{2.1\\_algorithm\\_svm.ipynb} - Implementation of SVM algorithm

625: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.2\_algorithm\_random\_forest.ipynb}{2.2\\_algorithm\\_random\\_forest.ipynb} - Implementation of Random Forest algorithm

627: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.3\_algorithm\_ridge.ipynb}{2.3\\_algorithm\\_ridge.ipynb} - Implementation of Ridge algorithm

629: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.4\_algorithm\_lasso.ipynb}{2.4\\_algorithm\\_lasso.ipynb} - Implementation of Lasso algorithm

631: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.5\_algorithm\_neural\_network\_tf.ipynb}{2.5\\_algorithm\\_neural\\_network\\_tf.ipynb} - Implementation of Neural Network algorithm

633: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/2.6\_algorithm\_xgboost.ipynb}{2.6\\_algorithm\\_xgboost.ipynb} - Implementation of XGBoost algorithm

635: \item Code: \href{https://github.com/bigdata-i523/hid306/blob/master/project/code/3\_ensemble\_kaggle\_submission.ipynb}{3\\_ensemble\\_kaggle\\_submission.ipynb} - Implementation of Ensembled algorithm

641: \item Input data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/train.csv}{train.csv} - Sample training dataset with housing attributes along with the sale price

642: \item input data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/test.csv}{test.csv} - Sample testing dataset similar to training dataset without the sale price

644: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_svm.csv}{kaggle\\_python\\_svm.csv} - Predicted Housing Sale Prices from SVM algorithm

646: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_random\_forest.csv}{kaggle\\_python\\_random\\_forest.csv} - Predicted Housing Sale Prices from Random Forest algorithm

648: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_ridge.csv}{kaggle\\_python\\_ridge.csv} - Predicted Housing Sale Prices from Ridge algorithm

650: \item Output data file: \href{https://github.com/bigdata-i523/hid}

306/blob/master/project/data/kaggle\_python\_xgboost.csv}{kaggle\\_python\\_xgboost.csv} - Predicted Housing Sale Prices from XGBoost algorithm  
652: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_lasso.csv}{kaggle\\_pythont\\_lasso.csv} - Predicted Housing Sale Prices from Lasso algorithm  
654: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_neural\_network.csv}{kaggle\\_python\\_neural\\_network.csv} - Predicted Housing Sale Prices from Neural Network algorithm  
656: \item Output data file: \href{https://github.com/bigdata-i523/hid306/blob/master/project/data/kaggle\_python\_ensemble.csv}{kaggle\\_python\\_ensemble.csv} - Predicted Housing Sale Prices from Ensembled algorithm

figures 27  
tables 2  
includegraphics 15  
labels 29  
refs 33  
floats 29

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 Applications in Real Estate Analysis

Elena Kirzhner

Indiana University Bloomington

3209 E 10th St

Bloomington, Indiana 47408

ekirzhne@iu.edu

## ABSTRACT

Big Data analysis reveals and comforts buyers with knowledge and facts about the neighborhood, its people and trends. Reducing risk of buying and predicting changes in home value for potential buyers.

## KEYWORDS

i523, hid320, Big Data Applications and Analytics, Real Estate

## 1 INTRODUCTION

When one mentions American dream, home ownership is first aspect that comes to mind. Another part of the American dream is financial success and wealth building. Buying your dream home to raise the family is obvious part of the real estate. For most Americans buying a home is the largest purchase they will ever make. Coupled with the fact that most conventional mortgages span 30 years research and analysis required to make educated choice should not be taken lightly as it will have implications on lifestyle for practically 40 percent of your lifetime. Successful investment in your home, potential rental property or land can lead to financially windfall. Failure to make right choices in real-estate purchases may have disastrous consequences. Financial ruin is obvious part of the equation. Majority of divorces in the united states are caused by financial duress in the households. Resulting in stress negatively affecting one's health.

The latest trend in real estate is application of Big Data. Big Data manipulation is booming and transforming the industry. We are seeing a huge move in usage of Big data and analytics. Companies build property matching online software based on customers behavior and their needs. The opportunities of Big Data are truly endless. It creates the power to change our thinking in decision making and develops efficient business approach by extracting variety of collected data points and reducing risks for consumers.

Big Data is already changing real estate industry by optimizing consumers search, offers recommendations on real estate websites to potential buyers and sellers. Utilizing Big Data in real estate could match customers with their desired home. It might include how many bedrooms they need, what neighborhood fits best, affordability, schools, crime rates, potential business property for rent, location and communities.

When using Big Data and analytics, it is possible to review patterns to understand whether the property is a good investment and a great match to potential customer. It is also possible to analyze what buyers are selecting more often and based on that data create a model.

When selecting a specific house for sale, Big Data integration within online websites made it possible to analyze local surroundings, sale patterns and neighborhood personality of each area. It created a knowledge comfort by having facts of the neighborhood, its people; and therefore reducing the risk of buying or investing in the wrong property.

## 2 BIG DATA IN REAL ESTATE BUSINESS

Risk mitigation is essential part of the way Big Data is transforming real estate. Open data across the internet and variety of Big Data tools added strong force for analysis in decision making of choosing right property or home. It equipped customers with the valuable information by extracting the data and cross analyzing it.

Big real estate agencies such as Realtor [13], Zillow [22] and Trulia [16], are pioneering those tools and provide estimated forecast of the property value from 1 to 10 years. Additionally, they provide information about the neighborhood trends, estimate mortgage payment, cost of ownership, history of the property and current value. The calculation is based on variety of public data records, market information, user data points [21] by using Big Data analysis formula developed in-house.

### 2.1 Real Estate Industry Evolution

Automated valuation methods have been used for a very long time. For decades banks utilized "Automated Valuation Model" to estimate home values. At one point banks wanted to exclusively rely on this model more than home values provided by professional appraisers. That practice led to problems with by omitting important nuances about condition resulting in overvaluation and undervaluation of properties. Big Data analysis and property estimates generated by online real estate giants are the next step in the evolution of real estate industry. This evolution diminishes importance and need for a real-estate agents as it is able to gather a lot of tribal information known only to experts in the area. That means this change can impact job market for over six hundred thousand active agents in the US.

### 2.2 Real Estate and Artificial Intelligence

Real estate businesses worry that unlocking the vast amount of data about properties could transform the business to be powered by artificial intelligence.

However, based on the GeekWire article [7] big data and artificial intelligence will not replace real estate agents. Robots are just big help and enrichment to the business. It created much better and safer decision making models. Artificial intelligence will help to deliver information about real estate transactions and trends to consumers. It says that in future Amazon voice or Siri could provide

useful information about popular housing trends and market value. Additionally, it can reveal the data on how many people were interested in the property and bids.

So far it is not a robot that is thinking and proactively making decision, it is just a voice based system that extracts the information from Big Data.

For the last twenty years, industry worries about loosing jobs in that area. However, the industry stayed the same. People still want an advice before making an important decision. Even thought, there so much more information and streamlined sales, individuals that want relationships, empathy and connecting with people are still there.

Obviously, there is some fear in real estate that robots can rock the world for real estate business. However, Big Data empowers agents with information and data, it is making them better providers with higher service.

### 2.3 Online Real Estate Agencies

Online real estate agencies calculate market value by using proprietary formulas. They are not providing expert estimation but a starting point in estimated property monetary worth. It is calculated from public data and surveys, by utilizing special features, market conditions and location. Additionally, they encourage consumers and homeowners to expand online data by doing other investigations such as comparing market prices for around areas, working with a real estate agent, getting an appraisal from an expert and visiting the house [21].

For example Zillow, developed a Zestimate prediction[21], which is Zillow's estimate of a home that currently on sale, one to ten years from now. The provided information based on current house and market condition. Other real estate agencies with online presence competing with Zillow, like Trulia and Realtor for example have developed similar proprietary formulas to assist customers.

Also, the companies provide rent estimates that would help evaluate potential monthly rental price by developed in-house algorithmic formula. Variations in rental prices can also happen because of different factors, additional investments, or length of lease.

Big Data information affects the forecasting. As an example, the amount of rental listings in a specific area affects how much we know about approximate prices in that area for condos, apartments and houses. Based on number of properties for rent, the prediction becomes more accurate. Homeowners can also update and provide information online about their needs or property, which helps even more for predicted accuracy.

The formula they use to estimate rent prices is comparing similar homes and apartments in the given area. Comparing bedrooms, square footage and other details. Then prices are being compared, and pattern to rental prices is shown.

Big data analysis provides unbiased information. Although, majority of real estate agents are esteemed professionals looking out for client's best interests they are still. There are still those that would like to manipulate client's opinion to benefit themselves. For example, if a particular house have been on the market too long and the agent might lose the listing there is a possibility that some shortcoming of the property will be omitted by the agent in order to complete the sale. Same can be said about agents trying

to achieve some sales goals or quotas. However, if the potential buyer conducts the research using Big Data all information will be available. Put simply, data does not lie.

## 3 REAL ESTATE ANALYSIS

Big Data is widely used by agents and real estate agencies to understand and improve how to target potential buyers. But the great thing about Big Data is that customers benefit from it as well. They can use free public resources with tons of data and information maps with different data analyzing tool options.

Latest tools allow to utilize Python to cross mix and match different values and data sets to analyze complex data. Prior to having these tools available such analysis would be an impossible task for individual users and required immense human and computing effort to complete. It is possible to visualize it by rendering correlations and trends. It reveals stunning insights in to chosen property for rent, business or home.

There is so much information that it is important to understand which data is relevant to consumers and improves decision making. It is useful to analyze the data-sets when considering investing [3]. The analysis can provide variety information and make the educated decision on the investment.

### 3.1 Big Data Tools

Analysis of these featured data points could be done with Python tool sets and libraries.

Python is a great programming language with variety of options. It is object oriented, semantically structured and great for scripting programs as well as connecting other programmable components. Python is considerably easy to learn and because of its high productivity and also became one of the favorite tools for programmers and data scientists. It contains libraries that are script importable and usable for a lot of use cases, such as image modification, scientific data analysis and server automation. Python world has been around for thirty years and a lot of code was written with multiple contributors. Variety of options built up on how to visualize the data [19].

The most common type of visualization is a simple bar chart and line graph [14]. It is popular and commonly used type of visualization to make comparison between values and variety of categories. It can be vertically or horizontally oriented by adjusting x and y axes, depending on what kind of information or categories the chart requires to present. Parameters need to be identified, such as axes, similarities, title and decided on what exactly the visualization supposed to show.

To make a simple bar chart, a number some of the most popular tools and libraries that have been invented for plotting the data could be utilized. These include the most used and common tools such as: Pandas, Seaborn, Bokeh, Pygal and Plotly.

Additionally, just like any other programming language issues, errors or questions with the libraries can be found on stack overflow page by Google search.

### 3.2 Data Analysis

For the purpose of this project, bar chart and graphs visualization methods with pandas modules in Python have been rendered and

explained. The simple form of this plot looks acceptable and easy to read.

The techniques were done within Jupiter notebook [9].

Jupiter notebook is great for running data sets analysis and for calculation projects. Jupiter notebook documents are readable files having the analysis description and the results in figures and tables as well as exportable files which can be executed to perform data analysis. It allows to render images and move values back and forth between different modules and coding languages.

The data-sets collected from clsearch.com [5], data.gov [17], zillow.com [22] and uploaded to the class's Google Drive to demonstrate the trends and patterns between each output.

The data includes both geographic and social data-sets evaluated by ratings in rows and titles in columns to keep it simple. The data set for both cities is being used for all examples that are demonstrated below. The point of the visualization is to understand the data in visual platform and make an informative decision based on rendered data.

A simple example of two properties in Tarzana, California versus Calabasas, were compared and exported for read.

Tarzana City is a wealthy neighborhood in the San Fernando Valley region of the city of Los Angeles, California. Tarzana was purchased in 1919 and developed on the site of local elites and named by Edgar Rice Burroughs, author of the popular Tarzan books. He established Tarzana and later sold it to local farmers [20].

Calabasas City located in the hills west of Malibu, in the San Fernando Valley region of the city of Los Angeles, California. The area established in 1991 and the name was derived from Spanish word "calabaza", meaning pumpkin. The legend has it that in 1824, a Mexican rancher spilled a wagon of pumpkin seeds and it spouted alongside the road. Therefore, the area was named Calabasas, the pumpkin land [20].

From a quick glance both areas are very similar and are located within 10 miles of each other. Both Tarzana and Calabasas are influential and desirable neighborhoods with lots of high priced homes. How does one differentiate between the two in order to find the right investment?

Big Data is the answer. Specifically in states like California. In California Big Data application benefits greatly from availability of public records such as sale price as apposed to certain "non-disclosure" states. There sale prices for homes are not disclosed in public records.

The analysis combines several main components, including property characteristics in the area, crime rate, quality of life, pollution, race and ethnicity, population growth, family household, house value, business field, employment, schools and future home value.

### 3.3 Property Characteristics

Big Data analytics can help in connecting needs of a buyer and providing neighborhood demographics. The quality of population in the neighborhood will influence who buys the house and who lives there. It is important to identify what is important to you and make sure those items are covered in the research. For example, if you are a student you will probably look for a densely populated

location around universities, closer to food locations and communities. Things like public transportation, nightlife, and bars will be very important to you and will be prioritized over other things. If you are married with kids, your best choice would be location with good schools and low crime. Parks, playgrounds and traffic and noise pollution around the house will be paramount. Most parents would love to find a nice quite cul-de-sac house. Young working professional would prefer to be right in the middle of things on a busy boulevard.

Latest Big Data collections made it all possible for real estate website to provide that information to potential buyers. Websites such as United States Zip-codes [2] collect information from public records and make it available in exportable format as well as for reviewing and analyzing local neighborhoods by states and zip codes input.

### 3.4 Crime Rate Indexes

Crime Big Data is available now and helps to see patterns and avoid areas with unfavorable statistics. The Los Angeles Police Department [1] already uses the data to show which areas in Los Angeles are hot-spots of crime.

Crime rates are being calculated by comparing the national levels of the average 100 [6]. For example, if score is 150 it means that it is 1.5 higher risk of crime than national average level. The data is coming from police department reports and public records. Additionally, the Federal Bureau of Investigation also provides factual information for ranking [18]. Furthermore, the research on crime can be extracted from United States Department of Justice via Uniform Crime Reporting Program [11].

In this example [9], running the crime data sets of Tarzana and Calabasas showed that Calabasas is in much better shape and safer place to live as compared to Tarzana, which is around the average of national rate. The total crime risk in Tarzana slightly higher than national average, it is 108, meanwhile Calabasas is almost 3 times lower, it is 24. The murder risk is 118 compared to 24 in Calabasas. Rape risk is 70 in Tarzana and 35 in Calabasas. Robbery risk in Tarzana is almost twice higher than in Calabasas, it is 125 versus 77. Assault risk three times higher in Tarzana, it is 127 versus 48 in Calabasas. Burglary risk twice higher in Tarzana as well, it is 55 versus 27. Larceny risk in Tarzana is overwhelmingly high, it is 73 versus 9 in Calabasas. Motor vehicle theft risk in Tarzana is 118 versus 39 in Calabasas. Based on these findings, it is defiantly safer to live in Calabasas [fig 1].

Based on these finding, it is defiantly safer to live in Calabasas as shown in Figure 1 [9].

[Figure 1 about here.]

### 3.5 Education Levels

Next run was done on educational level of residents. Big Data includes data of resident's education level and makes it possible to collect data about an individual resident and provides insightful information about social level interaction. The data extracted and combined from variety of sources including international school districts.

The education rating filtered by zip codes represents the percentage of people in the area who have attended colleges and received degrees. It does not represent performance and specific schools.

The rendered data showed [9] that residents in Calabasas are higher educated by 7 percent with Bachelor degrees and 6 percent higher with graduate degree, as shown in Figure 2 [9].

[Figure 2 about here.]

Based on the Economic Policy Institute study [4], there is a clear correlation between higher educated workforce and economic success within state and ability to grow. Additionally, higher educated people are good for state budgets, since workers with higher income contribute more through taxes.

### 3.6 Life Quality Standards

The next important consideration in buying a property, searching for a house and making a decision is quality of life standards in that area. Big Data and latest methods of data collection can lead to improvements in quality of life for residential areas. It can find neighborhoods that are safer, cleaner, more entertaining and a better place to live specifically tailored to potential buyer.

The data-set of life quality obtained from variety of sources, including public Google searches, social media and local study groups. The quality of life is being measured by how residents are being effected by crimes, weather, education, entertainment, religion, medical support and food supply. The positive decision variables calculated by amusement, education, culture, media, religion, weather and restaurants. The negative decision is based on the level of crime, natural disasters and mortality. The national level is being compared to 100 [5].

Rendered data showed that amusement index equal to 110 in Tarzana and 130 in Calabasas. What that means is that in Calabasas there are more community events and entertainment. Culture is 142 in Tarzana versus 129 in Calabasas. Culture refers to artistic development. Earthquake index 362 in Tarzana compared to 318 in Calabasas. this is a very interesting point considering that both neighborhoods are very close to each other. But since Tarzana's Earthquake index is higher associated insurance will likely be higher as well. Raising cost of ownership. Medical index is 137 in Tarzana and 116 in Calabasas. If you are working in the medical field this might be an important topic for you as it will help you find employment closer to home. Reduce your commute time, minimizing wasted time spent in California's infamous gridlock traffic. Mortality is much higher in Tarzana, it is 144 versus 95 in Calabasas. Religion is better in Tarzana, it is 154 compared to 96 in Calabasas. Religion refers to houses of worship and religious establishments. Restaurant index about the same, it is 144 in Tarzana and 139 in Calabasas. Weather is better in Tarzana, it is 16 versus 10 in Calabasas. That is another interesting observation considering that both neighborhoods are minutes away from each other.

Based on the data, overall quality of life is equal between two cities, as shown in Figure 3 [9].

[Figure 3 about here.]

### 3.7 Air Pollution

Big data can control and reveal pollution levels of particular area [8]. It is one of the main causes of health problems in the population and preventive cause death.

Over 80 percent of residents living in urban areas are vulnerable to poisoning from pollution. Cancer is one of the leading cause of deaths for both men and women; and exposure to pollution at early may have life-long negative consequences.

Monitored areas show that air quality levels exceed the safety levels [12]. Additionally, the World Health Organization warns that most populated states are most affected.

Government is aware of this problem, therefore collecting and monitoring the data regarding air quality has increased. The data is being shared between universities and air quality maps for further development. The data is openly shared and prepared for Big Data analysis.

Even though Big Data will not reduce the pollution by itself, it provides tools to visualize the problem which is especially helpful when choosing a place to live.

The exported data-sets showed [9] that carbon monoxide is extremely high in both cities. It is 186 in Tarzana and 183 in Calabasas. The national level is being compared to 100 [5]. Based on the data, overall air pollution index is about the same in both areas, as shown in Figure 4 [9].

[Figure 4 about here.]

### 3.8 Race and Ethnicity

Big Data can reveal a lot of information about population by using zip codes. It shows profiles of people who live there. Understanding ethnicity and identity of the community influence will help with decision.

The standard of maintaining, collecting and presenting federal data on race and ethnicity [10] were revised and improved on collecting quality about two decades ago. In accordance to best analysis practices, federal agencies conducting researches to better understand ethnic and race diversity.

The language to describe the ethnicity and race keeps changing to resonate with the category of residence and adding new meaning to not make it discriminatory. The general rule became that race and ethnicity should not be interpreted as being a science.

Based on the rendered graph, most population in Tarzana and Calabasas consist of white and non-Hispanic residents, as shown in Figure 5 [9].

[Figure 5 about here.]

That information provides insight about communities and relatedness to the buyer.

### 3.9 Population Growth

Leveraging Big Data in population growth might be helpful for economic growth prediction and future development.

For the recent centuries, population growth jumped dramatically [15]. How fast the population is growing can influence area homes and businesses development. Allowing for more business opportunities.

Educated people can contribute to the development with increased skills and knowledge. However, it is also important to look not only on the total population size, but also population growth rate.

Based on the data visualization, population size in Tarzana is higher by 3,000 residents than in Calabasas, as shown in Figure 6 [9].

[Figure 6 about here.]

Both in Calabasas and Tarzana, the rate was rapidly increasing from 1900-2000, and there was not much progress since then. Population density in Tarzana is 4,048 versus 856 in Tarzana. City area size in square miles is 7.44 and 31.67 in Calabasas. This information provides insight that Calabasas has much more opportunities for future growth and development. New housing and real estate development is achilles heel in California. State struggles to provide all existing residents with affordable housing. Compiled with population growth and migration of new residents the problem becomes even harder resolve. By having additional development space Calabasas growth potential is much higher compared to Tarzana.

### 3.10 Family Household

Big Data and internet of things are making its existence common place in each household. Only 15 years ago home computers were the only smart device in the house. Now even vacuums and thermostats are connected. Our homes are goldmines of data. Getting family household data summary instantly tells about the type of people in these areas and obtained knowledge can be used to help with buying decision.

Household definition refers to type of family and people living in a household structure. Household data is useful when consumer wants to know about the type of people living in that area and relativity.

Based on the combined data-set results [9], full family household is 64 percent in Tarzana and 76 percent in Calabasas. 48 percent are married in Tarzana, and 62 percent in Calabasas. Therefore for married families with kids it makes more sense to live in Calabasas.

### 3.11 Property Value

Big Data is being used to analyze property values. Real estate agencies, such as Zillow [22], estimate values based on Big Data collection tools and using their algorithm [21]. They combine information from variety of sources and provide insightful information to buyers, sellers or brokers.

Based on the data analysis [9], it shows that Calabasas prices are higher than in Tarzana by 23 percent. That insight shows that more financially able residence live in Calabasas.

To confirm that, the income data was calculated. Based on the rendered data as shown in Figure 7 [9], it proves that residence in Calabasas are more influential with higher income than in Tarzana.

The total income in Calabasas is higher by approximately 20 percent.

[Figure 7 about here.]

### 3.12 Employment and Occupation

The employment breakdown that derived from data, published by the Bureau of Labor Statistics showed that business field compared with employment field could help with predicting job opportunities.

Based on compared data sets, Health-care is leading employment field in Tarzana and Management in Calabasas, as shown in Figure 8 and Figure 9 [9].

[Figure 8 about here.]

[Figure 9 about here.]

### 3.13 Public Schools

Big Data in public schools are being used to fix education institutions and improve student scores and results. whereas in the past school performance was judged simply on average API scores of the students now student attributes data is further analyzed. This allows to identify subgroups of under-performing students. For example income levels of households are tracked to make sure that students from low-income families have the same opportunities to have better scores and grades as families from high-income families. It also provides tracking and comparison with schools in different districts. This helps school boards to allocate additional resources to schools that lack them. It also helps parents and home buyers identify schools and neighborhoods where their child could flourish academically.

The mined data could be used for decision making in property investment as well. Prospective buyers with kids are not only looking for good education and safe schools for their own kids, but also from stand-point of property value since homes located in good school districts are more desirable. The detailed information that can be found online made it easy to be properly informed. Compared data between two cities, showed that elementary schools have 38 percent higher rating in Calabasas, middle schools are 25 percent higher and high schools are the same. Schools in Calabasas are better based on these rating scores, as shown in Figure 10 [9].

[Figure 10 about here.]

### 3.14 Available Houses for Rent and Sale

Another shift in demographic preferences that has been observed is related to home ownership vs renting. Millennials are changing their spending habits when compared to previous generations. Food, health and entertainment take priorities over burdens expenses associated with home ownership. If that trend continues return on investment generated by buying rental properties will rise.

The best way to know if a house is a good investment is to check the rental properties near the area.

There is also a 1 percent rule of thumb to keep in mind. The rule is that a purchased home should be rented for 1 percent of the cost.

Based on the rental data, medium price in Tarzana 4,210 dollars per month, and Calabasas 4,085 dollars per month. It actually reveals that Tarzana rental properties are more expensive than Calabasas, even though the home prices in Calabasas are higher, as shown in Figure 11 [9].

[Figure 11 about here.]

Additionally, square footage was calculated. To get the price per square footage, the price of the area was divided by its square footage. The results showed that in Tarzana rent is slightly higher than in Calabasas, as shown in Figure 12 [9].

[Figure 12 about here.]

Therefore, it makes more sense to buy renting properties in Tarzana.

The lowest price of property in Tarzana is 700,000 us dollars, and in Calabasas it is 975,000 us dollars [9].

Based on the 1 percent rule, it does not make sense to buy and rent out in Tarzana or Calabasas.

### 3.15 Future Value

California housing is booming and crashing. Massive home equity destruction happened few years ago and reversed back.

When data-sets are analyzed, they can reveal insightful information and guide consumer decision making.

Based on the sales data was taken and generated, suggests that in spite of prices drops the value of houses goes up, as shown in Figure 13 and Figure 14 [9].

[Figure 13 about here.]

[Figure 14 about here.]

Calculated housing investment for the last 20 years had a growth rate of 5.46 percent [9]. By knowing a starting and ending value, it is possible to calculate the future value of an investment. Referencing the previous calculations [9], it predicts that house value will grow by 63 percent in the next 20 years.

## 4 CONCLUSION

Big Data potential to transform decision making in real-estate is immense. Home ownership is part of the American dream and Big Data will play a huge role in that process. It will allow potential buyers to have a better understanding of historic data and how it correlates to investment potential.

Big data will provide powerful insight to augment decision making process. Yet, it will not eliminate all risks associated with investment in real-estate. All risks must be evaluated and analyzed before buying and big data will provide plenty of tools for that.

Based on this analysis, it was determined that Tarzana and Calabasas properties are overpriced. Currently, renting is low compared to buying a property.

It is impossible to find properties in California that generate rents at around 1 percent of total property cost. You can not justify the prices and it is only for the privilege of living in San Fernando Valley region of the city of Los Angeles, California.

However, if you do still want to invest, Calabasas is a better choice for investing in a family home property and Tarzana for a rental property.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Juliette Zerick and Miao Jiang for their help, support and suggestions to write this paper.

## REFERENCES

- [1] 0 2017. *Federal Register The Daily Journal of the United States Government*. 0. <https://www.federalregister.gov>
- [2] 0 2017. *United States Zip Codes.org*. 0. <https://www.unitedstateszipcodes.org/91356/>
- [3] Andrew Beattie . 2017. *Top 10 Features of a Profitable Rental Property*. 0. <https://www.investopedia.com/articles/mortgages-real-estate/08/buy-rental-property.asp>
- [4] Noah Berger. 2013. *A Well-Educated Workforce Is Key to State Prosperity*. 0. <http://www.epi.org/publication/states-education-productivity-growth-foundations/>
- [5] CLRsearch.com. 2012. *Tarzana, CA 91356 Population Growth and Population Statistics*. 0. <https://www.clrsearch.com/Tarzana-Demographics/CA/91356/Population-Growth-and-Population-Statistics>
- [6] CLRsearch.org. 2012. *Community Demographic Information FAQ*. 0. <https://www.clrsearch.com/demographics/Demographic.Information.jsp>
- [7] John Cook. 2017. *Robots in real estate?* 0. <https://www.geekwire.com/2017/robots-real-estate-theres-nothing-see-zillow-co-founder-says-agent-jobs-safe/>
- [8] Arantxa Herranz. 2017. *Big data will control pollution in your city*. 0. <http://blog.ferrovial.com/en/2017/04/big-data-pollution-control-in-cities/>
- [9] Elena Kirzhner. 2017. *Big Data Applications in Real Estate*. 0. <https://github.com/bidata-i523/hid320/blob/master/project/project.md>
- [10] Management and Budget Office. 2016. *Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity*. 0. <https://www.federalregister.gov/documents/2016/09/30/2016-23672/standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity>
- [11] FBI/US Crime Statistics Management. 2017. *Uniform Crime Reporting Statistics: Their Proper Use*. 0. <https://ucr.fbi.gov/ucr-statistics-their-proper-use>
- [12] World Health Organization. 2016. *Air pollution levels rising in many of the world's poorest cities*. 0. <http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/en/>
- [13] Realtor.com. 2017. *Realtor.com - resource for home buyers and sellers*. 0. <https://www.realtor.com>
- [14] Naomi B Robbins. 2012. *Creating more effective graphs*. Wiley, 0.
- [15] Max Roser and Esteban Ortiz-Ospina. 2017. *World Population Growth*. 0. [https://ourworldindata.org/world-population-growth/](https://ourworldindata.org/world-population-growth)
- [16] Trulia.com. 2017. *Trulia is a mobile and online real estate resource*. 0. <https://www.trulia.com>
- [17] U.S. General Services Administration, Technology Transformation Service. 2017. *Real Estate Sale History*. 0. <https://www.data.gov>
- [18] Mark van Rijmenam. 2017. *The Los Angeles Police Department Is Predicting and Fighting Crime With Big Data*. 0. <https://datafloq.com/read/los-angeles-police-department-predicts-fights-crim/279>
- [19] Guido Van Rossum and Fred L Drake. 2011. *The python language reference manual*. Network Theory Ltd., 0.
- [20] Wikipedia. 2017. *Tarzana, Los Angeles*. 0. [https://en.wikipedia.org/wiki/Tarzana,\\_Los\\_Angeles](https://en.wikipedia.org/wiki/Tarzana,_Los_Angeles)
- [21] Zillow.com. 2017. *The Zestimate home valuation*. 0. <https://www.zillow.com/zestimate/#what>
- [22] Zillow.com. 2017. *Zillow is the leading real estate and rental marketplace*. 0. <https://www.zillow.com>

#### LIST OF FIGURES

1	Crime rate in Tarzana, CA compared to Calabasas, CA (100 = National Average) [9].	8
2	Educational percentage of people in Tarzana, CA compared to Calabasas, CA (Population Age 25+) [9].	9
3	Life quality of people in Tarzana, CA compared to Calabasas, CA [9].	10
4	Air Pollution Indexes in Tarzana, CA compared to Calabasas, CA [9].	11
5	2012 Population by Race and Ethnicity in Tarzana, CA compared to Calabasas, CA [9].	12
6	Population change since 1990 in Tarzana, CA compared to Calabasas, CA [9].	13
7	Income in Tarzana, CA compared to Calabasas, CA [9].	14
8	Employment field in Tarzana, CA compared to Calabasas, CA [9].	14
9	Business fields in Tarzana, CA compared to Calabasas, CA [9].	15
10	Public schools in Tarzana, CA compared to Calabasas, CA [9].	15
11	Houses for rent in Tarzana, CA compared to Calabasas, CA (3bd+ House For Rent (1,500-2,500 Sqft)) [9].	16
12	Price per sqft for rent in Tarzana, CA compared to Calabasas, CA [9].	17
13	Prices Growth Index in California [9].	18
14	Prices Growth Index in California [9].	19

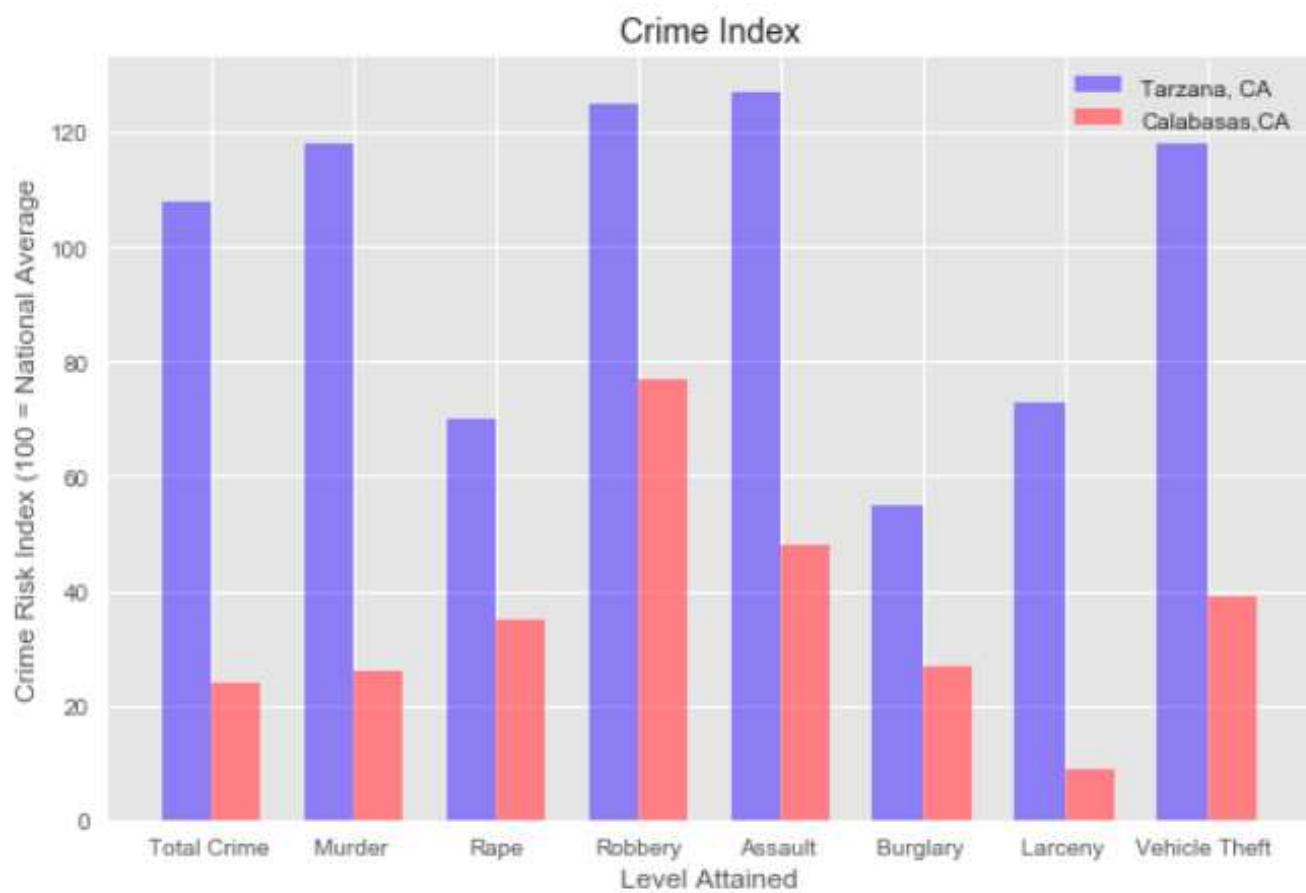


Figure 1: Crime rate in Tarzana, CA compared to Calabasas, CA (100 = National Average) [9].

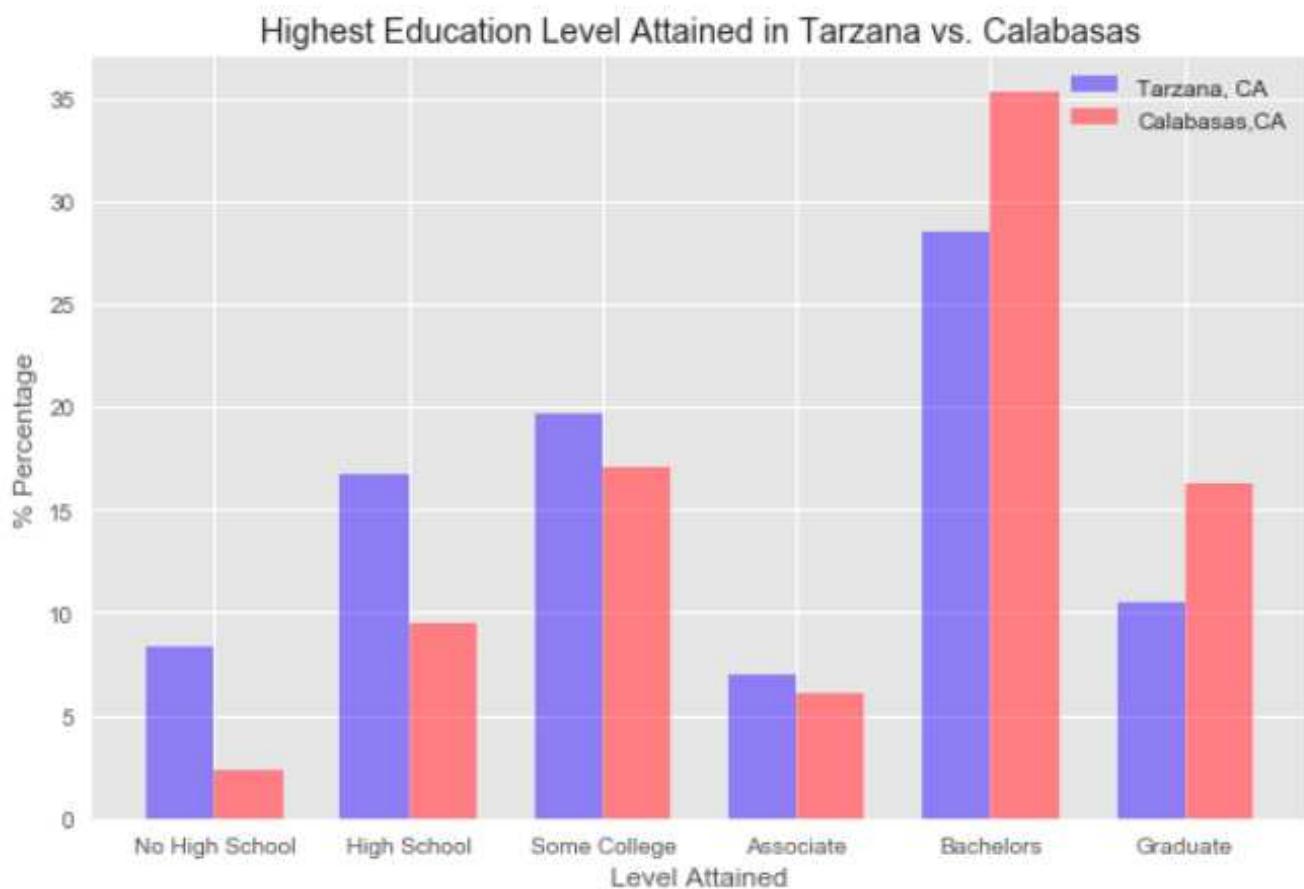


Figure 2: Educational percentage of people in Tarzana, CA compared to Calabasas, CA (Population Age 25+) [9].

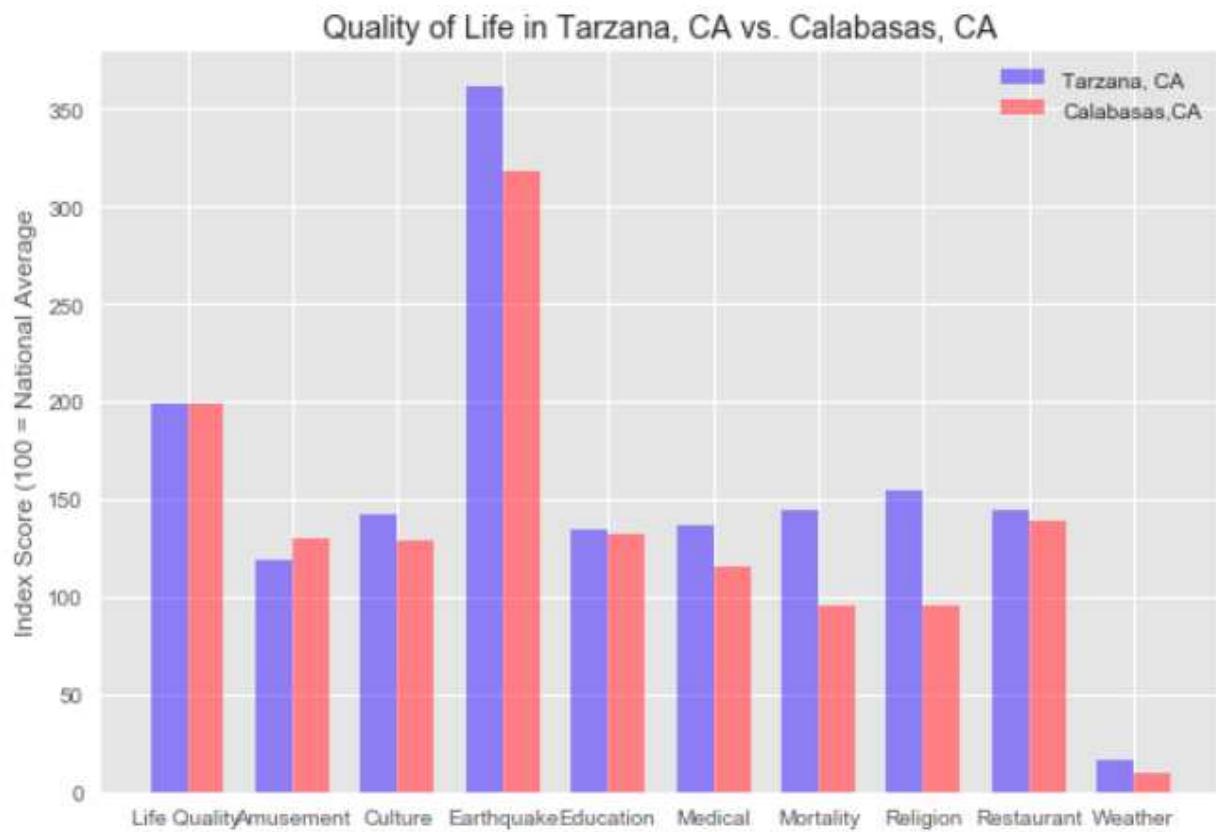


Figure 3: Life quality of people in Tarzana, CA compared to Calabasas, CA [9].

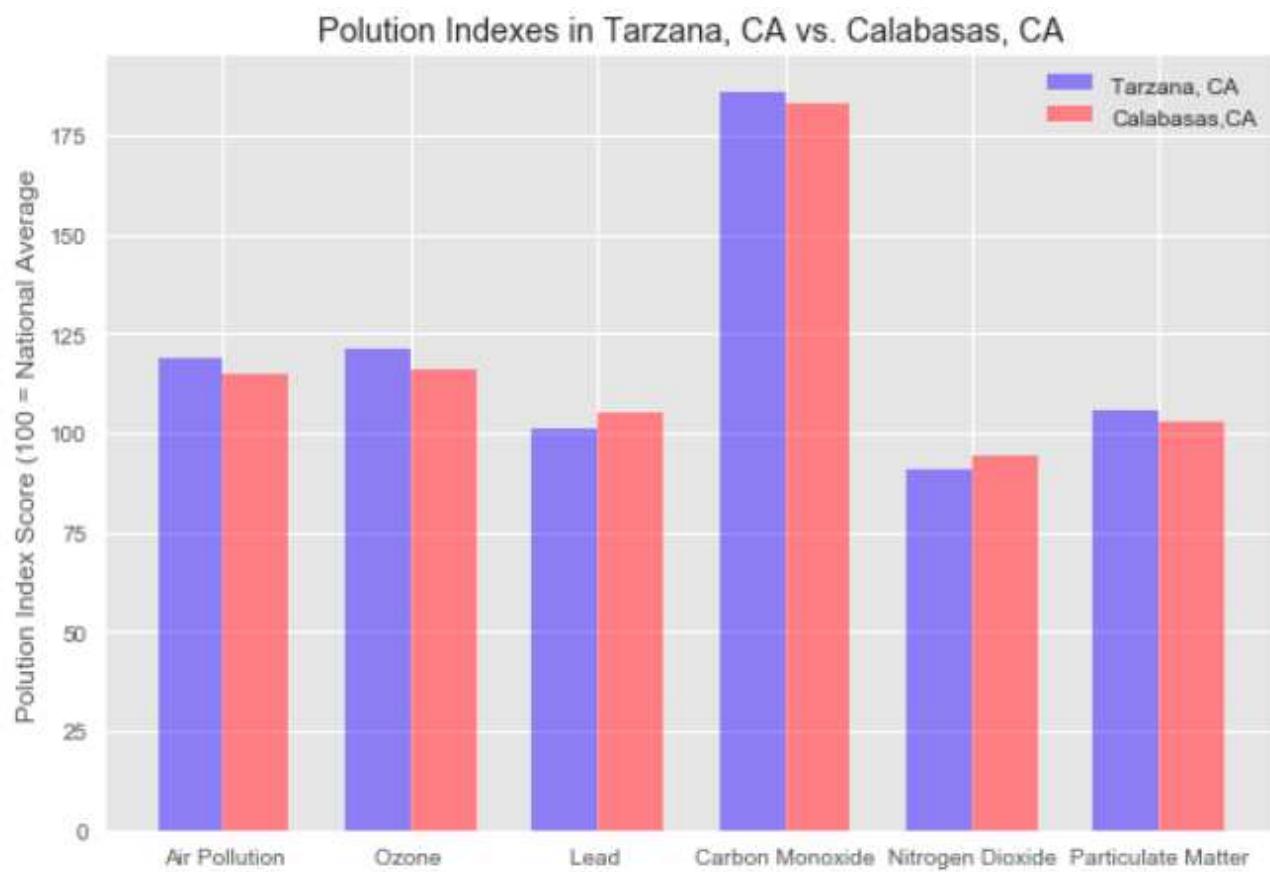


Figure 4: Air Pollution Indexes in Tarzana, CA compared to Calabasas, CA [9].

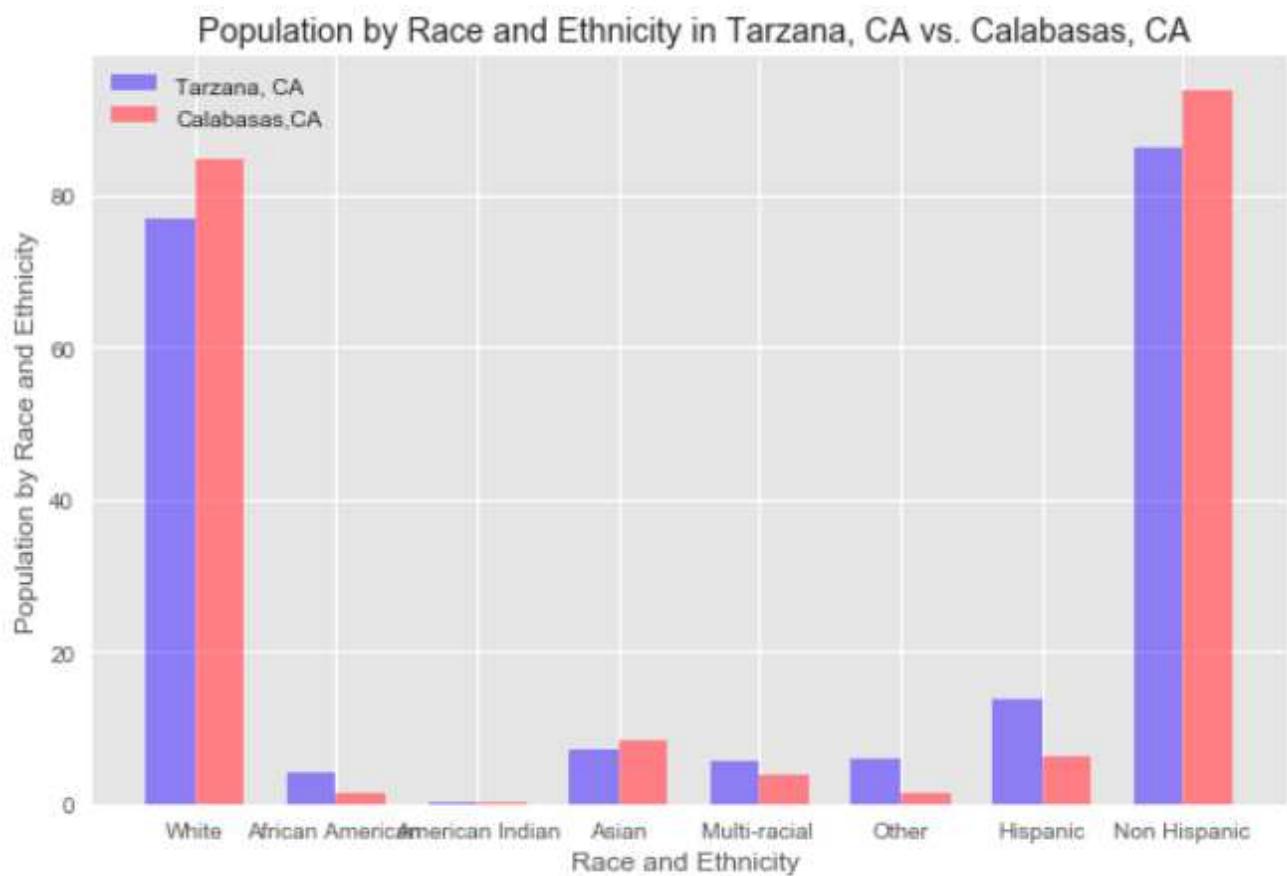


Figure 5: 2012 Population by Race and Ethnicity in Tarzana, CA compared to Calabasas, CA [9].

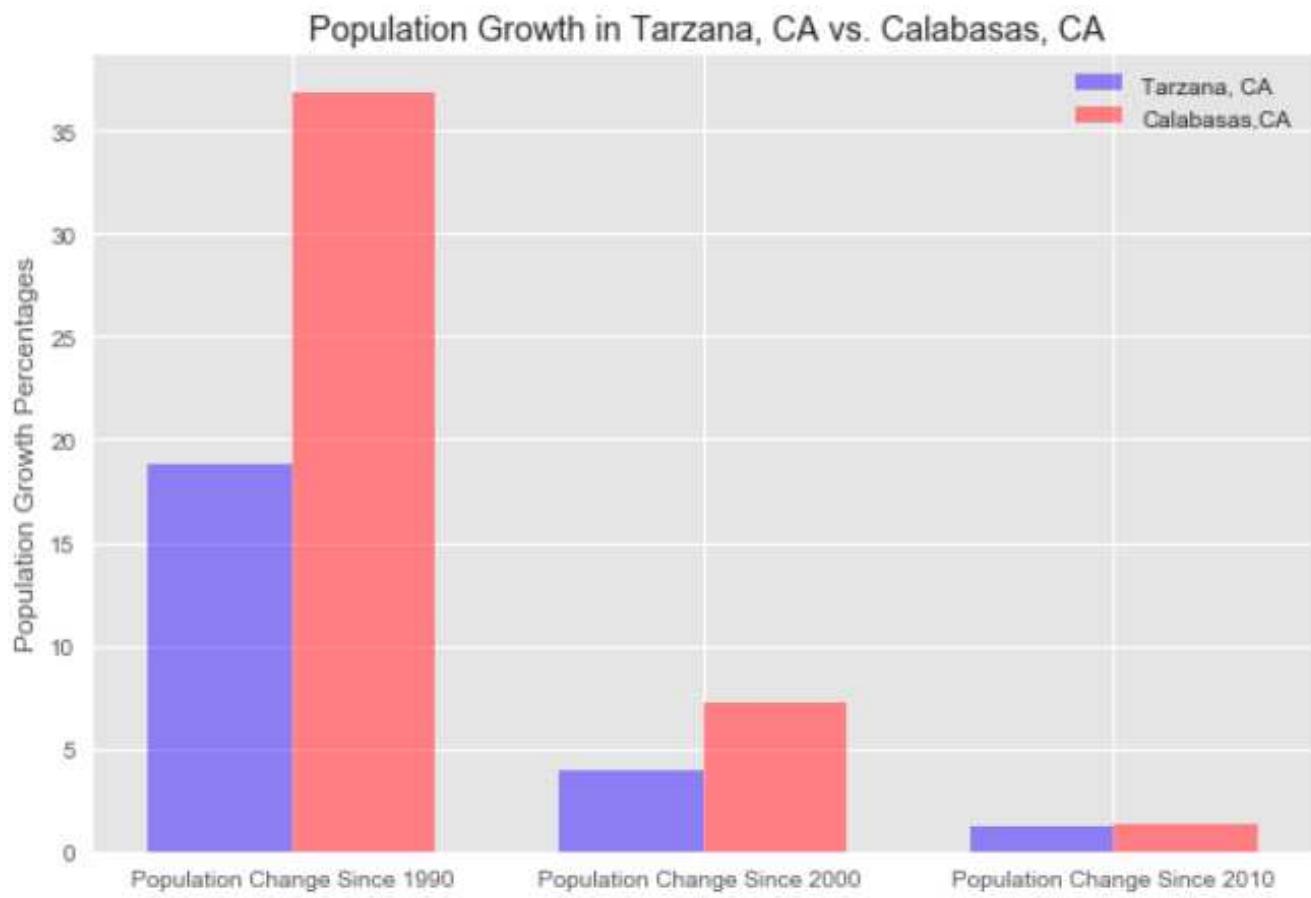
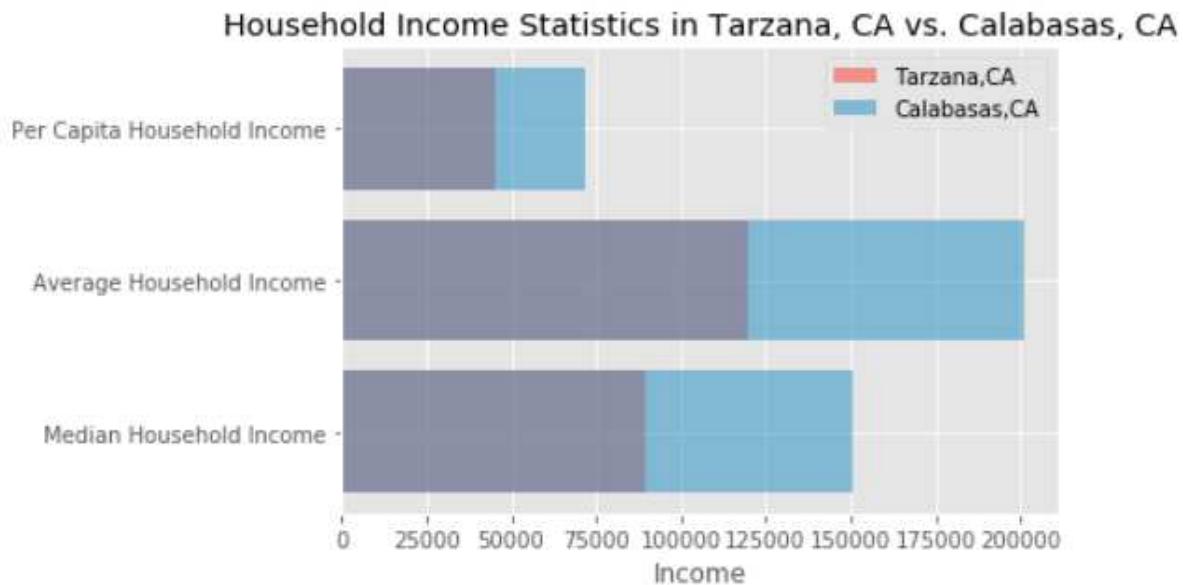
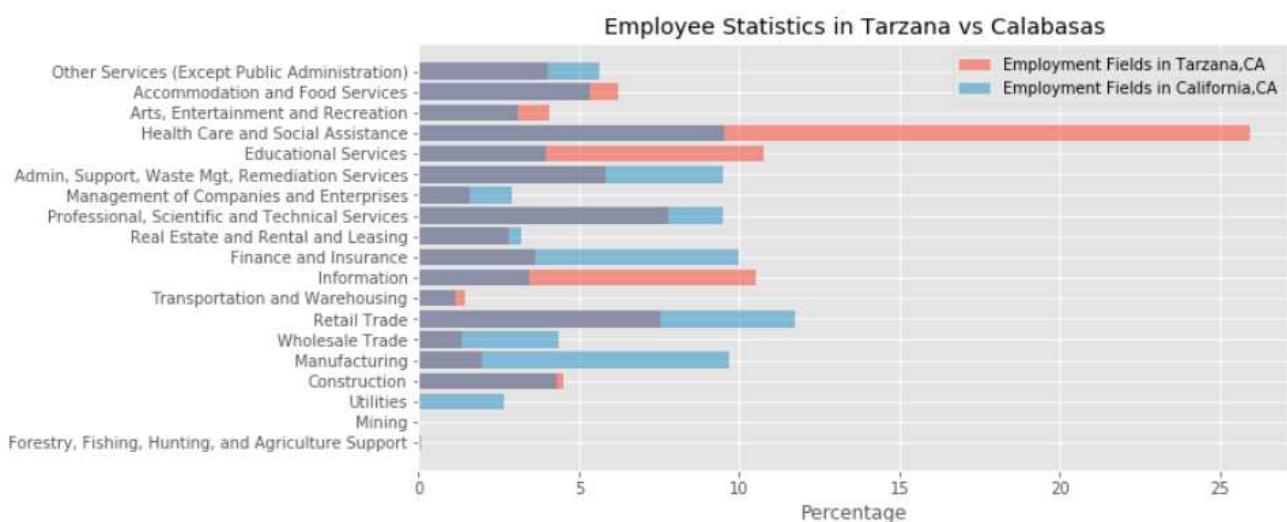


Figure 6: Population change since 1990 in Tarzana, CA compared to Calabasas, CA [9].



**Figure 7: Income in Tarzana, CA compared to Calabasas, CA [9].**



**Figure 8: Employment field in Tarzana, CA compared to Calabasas, CA [9].**

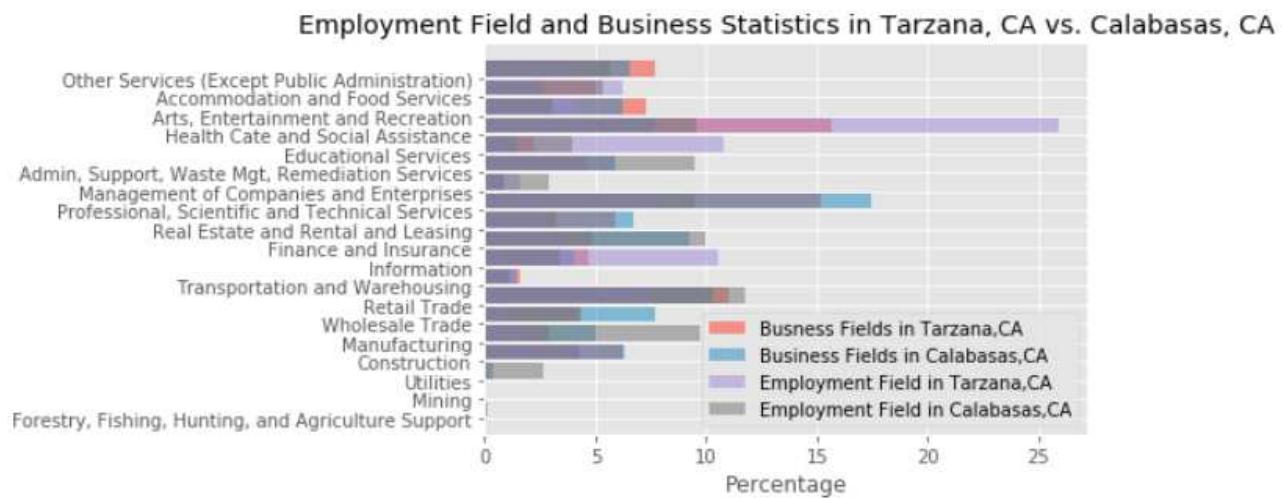


Figure 9: Business fields in Tarzana, CA compared to Calabasas, CA [9].

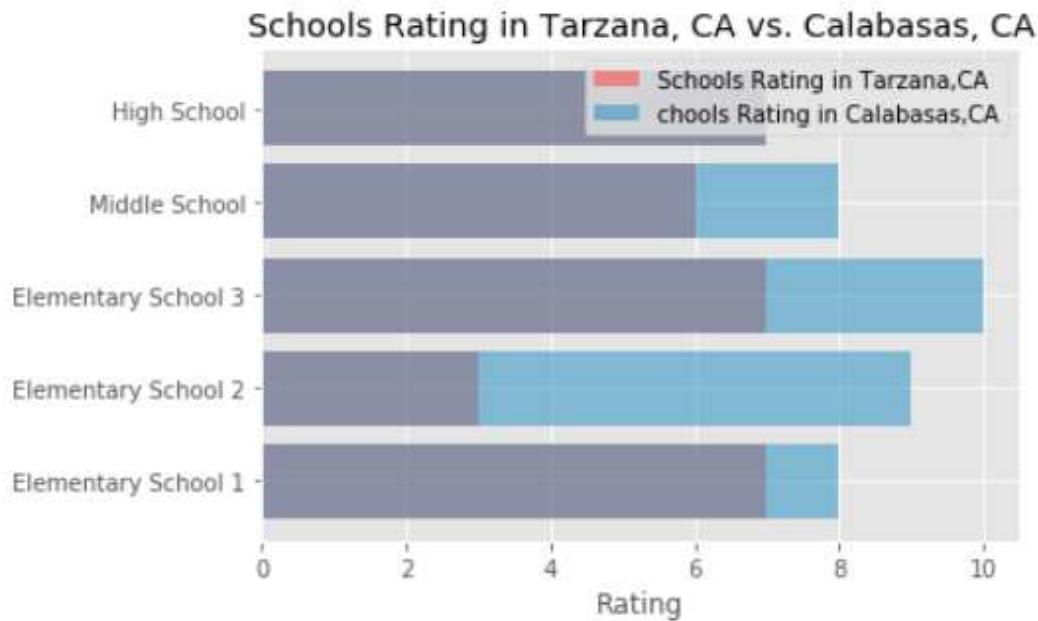


Figure 10: Public schools in Tarzana, CA compared to Calabasas, CA [9].



Figure 11: Houses for rent in Tarzana, CA compared to Calabasas, CA (3bd+ House For Rent (1,500-2,500 Sqft)) [9].

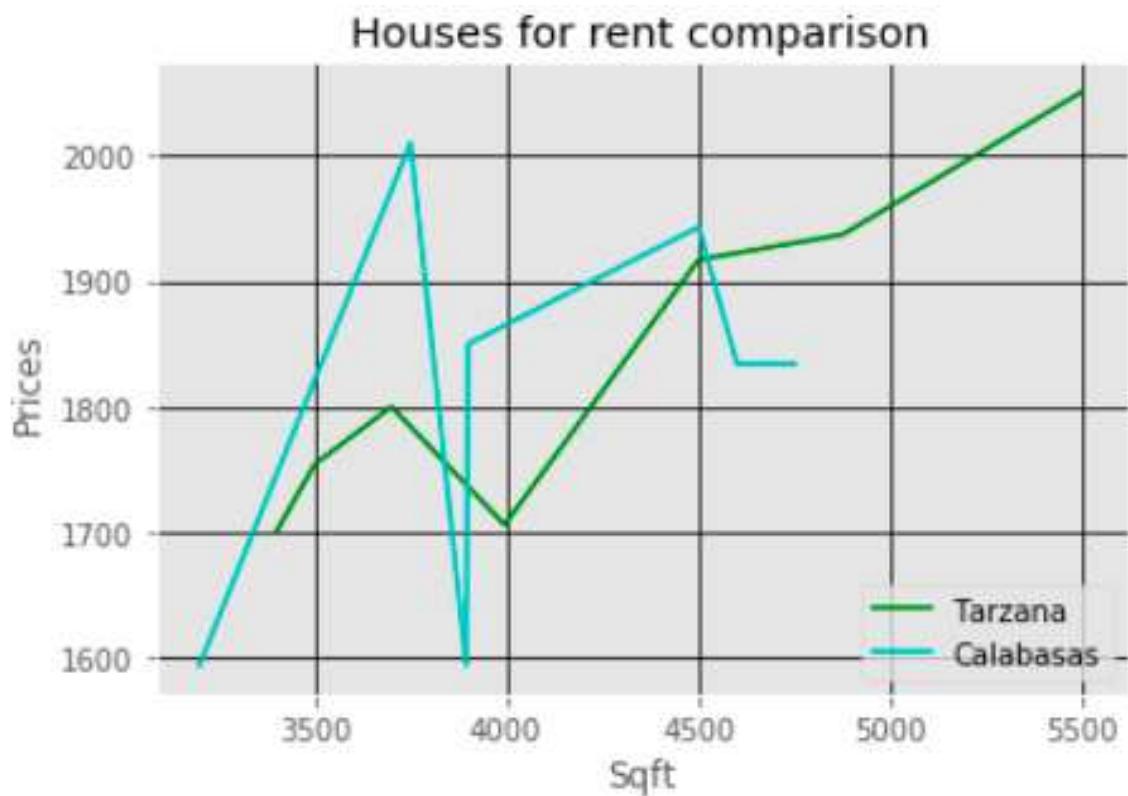


Figure 12: Price per sqft for rent in Tarzana, CA compared to Calabasas, CA [9].



Figure 13: Prices Growth Index in California [9].



Figure 14: Prices Growth Index in California [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
```

```
=====
report.bib:105:    @manual{education,
report.bib:145:    @manual{robots,
report.bib:153:    @book{van2011python,
report.bib:169:    @book{robbins2012creating,
report.bib:57:    @manual{clr,
report.bib:65:    @manual{zipcodes,
report.bib:73:    @manual{federal,
report.bib:81:    @manual{crime,
report.bib:89:    @manual{unicrime,
report.bib:97:    @manual{crime2,
```

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

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

```
[2017-12-10 13.51.20] 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 1.4s.
```

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

```
name: Elena Kirzhner  
hid: 320  
paper1: 100% Oct 31 2017  
paper2: 100% Nov 6 2017  
project: 100% Dec 03 2017
```

```
yamlcheck
```

```
wordcount
```

```
(null)  
wc 320 project (null) 5263 report.tex  
wc 320 project (null) 5567 report.pdf  
wc 320 project (null) 524 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
```

```
passed: True
```

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

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

```
passed: True
```

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

passed: False

floats

---

141: Based on these finding, it is defiantly safer to live in  
Calabasas as shown in Figure \ref{fig:figure1} \cite{md}.  
143: \begin{figure}  
145: \includegraphics[width=1.0\columnwidth]{images/figure1.png}  
146: \caption{Crime rate in Tarzana, CA compared to Calabasas, CA (100  
= National Average) \cite{md}.} \label{fig:figure1}  
156: The rendered data showed \cite{md} that residents in Calabasas  
are higher educated by 7 percent with Bachelor degrees and 6  
percent higher with graduate degree, as shown in Figure  
\ref{fig:figure2} \cite{md}.  
158: \begin{figure}  
160: \includegraphics[width=1.0\columnwidth]{images/figure2.png}  
161: \caption{Educational percentage of people in Tarzana, CA compared  
to Calabasas, CA (Population Age 25+) \cite{md}.}  
\\ \label{fig:figure2}  
175: Based on the data, overall quality of life is equal between two  
cities, as shown in Figure \ref{fig:figure3} \cite{md}.  
177: \begin{figure}  
179: \includegraphics[width=1.0\columnwidth]{images/figure3.png}  
180: \caption{Life quality of people in Tarzana, CA compared to  
Calabasas, CA \cite{md}.} \label{fig:figure3}  
195: The exported data-sets showed \cite{md} that carbon monoxide is  
extremely high in both cities. It is 186 in Tarzana and 183 in  
Calabasas. The national level is being compared to 100  
\cite{clr}. Based on the data, overall air pollution index is  
about the same in both areas, as shown in Figure  
\ref{fig:figure4} \cite{md}.  
197: \begin{figure}  
199: \includegraphics[width=1.0\columnwidth]{images/figure4.png}  
200: \caption{Air Pollution Indexes in Tarzana, CA compared to  
Calabasas, CA \cite{md}.} \label{fig:figure4}  
212: Based on the rendered graph, most population in Tarzana and  
Calabasas consist of white and non-Hispanic residents, as shown  
in Figure \ref{fig:figure5} \cite{md}.  
214: \begin{figure}  
216: \includegraphics[width=1.0\columnwidth]{images/figure5.png}  
217: \caption{2012 Population by Race and Ethnicity in Tarzana, CA  
compared to Calabasas, CA \cite{md}.} \label{fig:figure5}  
230: Based on the data visualization, population size in Tarzana is  
higher by 3,000 residents than in Calabasas, as shown in Figure

```

    \ref{fig:figure6} \cite{md}.
232: \begin{figure}
234: \includegraphics[width=1.0\columnwidth]{images/figure6.png}
235: \caption{Population change since 1990 in Tarzana, CA compared to
Calabasas, CA \cite{md}.} \label{fig:figure6}
256: To confirm that, the income data was calculated. Based on the
rendered data as shown in Figure \ref{fig:figure7} \cite{md}, it
proves that residence in Calabasas are more influential with
higher income than in Tarzana.
260: \begin{figure}
262: \includegraphics[width=1.0\columnwidth]{images/figure7.png}
263: \caption{Income in Tarzana, CA compared to Calabasas, CA
\cite{md}.} \label{fig:figure7}
270: Based on compared data sets, Health-care is leading employment
field in Tarzana and Management in Calabasas, as shown in Figure
\ref{fig:figure8} and Figure \ref{fig:figure9} \cite{md}.
272: \begin{figure}
274: \includegraphics[width=1.0\columnwidth]{images/figure8.png}
275: \caption{Employment field in Tarzana, CA compared to Calabasas,
CA \cite{md}.} \label{fig:figure8}
278: \begin{figure}
280: \includegraphics[width=1.0\columnwidth]{images/figure9.png}
281: \caption{Business fields in Tarzana, CA compared to Calabasas, CA
\cite{md}.} \label{fig:figure9}
290: Schools in Calabasas are better based on these rating scores, as
shown in Figure \ref{fig:figure10} \cite{md}.
292: \begin{figure}
294: \includegraphics[width=1.0\columnwidth]{images/figure10.png}
295: \caption{Public schools in Tarzana, CA compared to Calabasas, CA
\cite{md}.} \label{fig:figure10}
306: Based on the rental data, medium price in Tarzana 4,210 dollars
per month, and Calabasas 4,085 dollars per month. It actually
reveals that Tarzana rental properties are more expensive than
Calabasas, even though the home prices in Calabasas are higher,
as shown in Figure \ref{fig:figure11} \cite{md}.
308: \begin{figure}
310: \includegraphics[width=1.0\columnwidth]{images/figure11.png}
311: \caption{Houses for rent in Tarzana, CA compared to Calabasas, CA
(3bd+ House For Rent (1,500-2,500 Sqft)) \cite{md}.}
\label{fig:figure11}
314: Additionally, square footage was calculated. To get the price per
square footage, the price of the area was divided by its square
footage. The results showed that in Tarzana rent is slightly
higher than in Calabasas, as shown in Figure \ref{fig:figure12}
\cite{md}.
316: \begin{figure}

```

```
318: \includegraphics[width=1.0\columnwidth]{images/figure12.png}
319: \caption{Price per sqft for rent in Tarzana, CA compared to
Calabasas, CA \cite{md}.} \label{fig:figure12}
334: Based on the sales data was taken and generated, suggests that in
spite of prices drops the value of houses goes up, as shown in
Figure \ref{fig:figure13} and Figure \ref{fig:figure14}
\cite{md}.
336: \begin{figure}
338: \includegraphics[width=1.0\columnwidth]{images/figure13.png}
339: \caption{Prices Growth Index in California \cite{md}.}
\label{fig:figure13}
342: \begin{figure}
344: \includegraphics[width=1.0\columnwidth]{images/figure14.png}
345: \caption{Prices Growth Index in California \cite{md}.}
\label{fig:figure14}
```

figures 14  
tables 0  
includegraphics 14  
labels 14  
refs 12  
floats 14

```
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 ""
```

---

```
68: author = "",
```

```
76: author = "",
```

```
passed: False
```

```
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
```

# Big Data Analytics in Identifying Factors Affecting Bitcoin

Ashok Kuppuraj

Indiana University

Bloomington, Indiana 43017-6221

akuppura@iu.edu

## ABSTRACT

Pricing of Blockchain based cryptocurrencies are like a black box, as per theory the pricing compared to U.S dollar is based on a number of transactions however lot other factors like Dollar price, social media, Online threats supersede the transaction count. Big data and Analytics helps to identify the metrics impacting this variation and identify the correlation between them.

## KEYWORDS

i523, hid324, Big data, Predictive analytics, Random Forest, correlation, Blockchain, Bitcoin, Ethereum

## 1 INTRODUCTION

The start of the 21st century witnessed the evolution of various disruptive technologies, right from Big data, IoT, VR to Blockchain. When it comes to the blockchain, the sole winner is Bitcoin, with the growth rate of over 1327 percent [4], Bitcoin is disrupting the way banking system works. As the Bitcoin grows the acceptance and adoption grow along with that. Similar to any other currency in the world, Bitcoin's price deviates widely towards the positive side which created the opportunity for investment in it. Even though the same is not widely accepted everywhere, there is a grace to own Bitcoin citing its growth rate. Though the transaction counts haven't grown up, the retention of the coin has grown up making it a Digital Gold [15].

## 2 BITCOIN

Bitcoin is a progressed cryptographic cash and shared ledger that is completely decentralized, which implies it relies upon peer-to-peer trades with no bureaucratic oversight. Trades and liquidity inside the framework are somewhat based on cryptography. The concept was first introduced in 2009 [7] and is at this moment a prospering open-source gathering and portion sort out. In perspective of the uniqueness of Bitcoin's tradition and its creating choice, the Bitcoin is grabbing stacks of thought from associations, clients, and monetary experts alike. Specifically, for this technology to thrive, we need to recreate budgetary organizations and things that starting at now exist in our traditional, fiat cash world, make them available and specially fitted to Bitcoin, and other rising computerized types of cash. In technical terms, Bitcoin's is a shared ledger or a database running by a set of clusters, as the clustering is involved, a competition is set for the individual machines to acquire and update the ledger. The competition is in terms of hashing problem. The hashing needs multiple GPU's to perform validations and update the ledger. This competition eliminates the slower machines to be part of the network and improve the infrastructure's capacity, only by winning the competition a machine can be awarded some

Bitcoin as an incentive. Since one machine cannot process the competition problem, a set of peers come together to form a Mining pool and share their capacity and the incentives. We can gather useful mining statistics information from these mining pools.

## 3 PRICE PREDICTION

The Bitcoin market's cash-related basic is, clearly, a securities trade. To support money related to reward, the stock market prediction has turned out to be known ground which can be reused with the presence of high-repeat, low-dormancy trading hardware joined with solid machine learning figurings. Henceforth, it looks good that this desire is imitated in the domain of Bitcoin, as the framework expands more conspicuous liquidity and more people develop an excitement for placing profitably in the structure. To do accordingly, it is essential to utilize machine learning advancement to foresee the cost of Bitcoin [9].

### 3.1 Data Source

As Bitcoin is a decentralized and a transparent system, all the source of data can be gathered from the peer-to-peer networks. This peer-to-peer network is called as Bitcoin-mining pool [1]. The rate of block creation is adjusted every 2016 blocks to aim for a constant two week adjustment period (equivalent to 6 per hour.) The number of Bitcoins generated per block is set to decrease geometrically, with a 50 reduction every 210,000 blocks, or approximately four years. The result is that the number of bitcoins in existence is not expected to exceed 21 million [5]. The true source of data for Bitcoin analysis would be from Bitcoin mining pool. Coinbase is one of the main members of bitcoin pool from which we can gather mining statistics. In the process of identifying the features impacting Bitcoin's price fluctuations, not only the transaction volume impacts, even the popularity and people's trend towards it impact the price of the coins. Hence, data from Google is also gathered. As a currency's price also been altered by its exchange, supply, and demand, Ethereum's price data and transactional data is also acquired from Ethereum's exchange point. With all these data sources, we analyze the features impacting the Bitcoin's market price.

### 3.2 Feature Selection

Feature selection is one of the vital steps in any meaningful analysis of an expected outcome. A set of features have been selected to analyze its interdependence with Bitcoin's evaluation. The features are selected based on three wide areas, the first is Bitcoin mining data, second is social data and the last one is exchange data. The internal activities in the Bitcoin's infrastructure definitely reflect the changes or the fluctuations in the Bitcoin's network, Bitcoin's mining data is gathered from Coinbase. This is extracted from the

web service API provided by Quandl.com [8]. By making a REST call, CSV files containing the historical data is downloaded and processed. The second is the social data, which is extracted as a static data from Google trends [6], the main reason behind this data is when the popularity grows people tend to know or show interest in being part of the growth. With the impressive growth of more than 1000 percent in a year, this is considered as an important data. The last one is the exchange data, as a currencies price is directly proportional to the supply and demand, the supply of the currency can be impacted by the exchange to other currencies or commodity [16]. Ethereum is known to show a similar pattern in terms of growth and deviations [2]. Hence, there's price in US dollars and transaction volume is considered one of the features.

## 4 BIG DATA IN FEATURE ANALYSIS AND ALGORITHM'S EXECUTION

Feature extraction, transformation, and prediction can be synonymous with a conventional ETL methodology. Though few of the extraction is handled manually and the volume is comparably low, it is assumed that the data volume will be increased by modifying the extraction to real-time systems. When the extraction systems are changed, our code must be able to handle streaming data which can be related to "variety and volume " of the data. The next step is validating the data for anomalies, data miss and cleanse the data of issues which is synonymous with data cleansing. The later one is data processing, which includes data processing with multiple iterations and permutation consuming a lot of memory and other resources. These processing needs lead us in adopting Big data technologies in the entire lifecycle of the implementation. Apache Spark framework is identified as the end-to-end processing environment which is pre-loaded with redundancy, fault tolerance, in-memory processing, parallel processing, streaming, and Machine learning modules.

### 4.1 Execution with Apache Spark

"Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs". It also provides extensive support to Machine learning libraries(MLlib) and to streaming through Spark Streaming. The in-memory processing is implemented with the help of Resilient distributed dataset (RDD) [10].

[Figure 1 about here.]

Spark's architectures given in the figure 1 provide a glimpse of how different system in Spark is interfaced. The first level of interfacing to Spark is with high-level languages like Scala, Java, Python and R. Users implement their functionalities in these high-level languages. The primary executing components in Spark are Driver and Executor modules. The driver is the entry point for any implementation, the written programs will be executed in the main function of the Driver module, later converted to set of Directed Acyclic graph by the Spark APIs. DAGs are then executed in executors in the data nodes based on the data placement policy of the infrastructure. Four modules built on spark for serving the user's needs are SparkSQL, Spark Streaming, MLlib, and GraphX. Spark SQL and Machine Learning libraries(MLlib) are consumed in our

implementation and the future improvement would be on Spark Streaming which is used for Streaming requirements. SparkSQL and MLlib modules contain the implementation for DataFrames, SQL functionalities, and Machine learning libraries. The next level of the modules is data abstraction layer. Spark's basic data abstraction is Resilient Distributed Dataset (RDD), which is a fault tolerant partitioned data encapsulation datatype. The RDDs are lazily evaluated, hence a Directed Acyclic Graph is implemented to persist the state of the RDDs at each stage. With RDD, Spark can execute the transformation in parallel with fault tolerance. This implementation widely differentiates from conventional Python implementation which lacks this advanced logic. Apache's Spark 2.2 is used to implement all the ETL functionality. Spark is installed in the local system along with Anacondas, so Spark libraries can be consumed inside Python shell. To consume and process Pyspark libraries, sparkcontext is created which initiates the driver program. The spark context is bootstrapped with SQL and Spark session libraries so that Spark RDD and Data frames could be accessed under a single window.

As the abstract describes the necessity of the features impacting Bitcoin's price, the best metric to identify the relation between Bitcoin's price and its features is by identifying the correlation matrix provided by Charles Spearman. Spearman's function describes the relationship between two variable using a monotonic function [14]. Apart from identifying the correlation, these features can be modeled to predict the value of the dependent variable which is Bitcoin's value. The algorithm consumed for the predictions are Random forest and gradient boosted regression the Machine learning modules of Spark.

## 5 ARCHITECTURE

The architecture flow consists of three levels of components, first one is the Data extraction, second is processing and the final is visualization. The Figure 2 describes how the implementation is fitted over Spark's architecture. The logical implementation starts with extracting the data from the source and loading it over RDDs. With RDDs on the base, source data is validated for data miss and anomalies. With RDDs, all validation happens in parallel irrespective of any volume or variety of data. As RDDs are hash partitioned by default, it can consume any volume or type of data with consistent efficiency. Upon loading into RDDs, it is transformed to named columns as Dataframes which are indexed and more efficient in processing structured data. Pyspark dataframe is selected to increase the performance of the data processing even though Pyspark dataframe API is not equipped with rich functionalities similar to Pandas dataframe and Pyspark dataframe can execute the transformations in parallel whereas Pandas cannot. Machine learning algorithms are implemented over the Dataframes and generated model is executed and persisted as array objects for visualization.

[Figure 2 about here.]

### 5.1 Technologies

Technology stacks used in our implementation are,

- Python 2.7
- Pyspark 2.2
- Jupyter 5.0.0

## 5.2 Data Extraction

The data is sourced from Quandl.com, a public data service for various types of data, Bitcoin's mining data from 2015 till current date from Coinbase's mining pool. The data is in CSV format with Bitcoin's transaction details and its corresponding date associated with it.

The second set of data is from Etherscan, an open source portal for Ethereum transaction details, from which the transaction count and the price in US dollars are extracted. The third dataset is about the people's trends on Bitcoin's popularity from Google, the granularity of this data is on weekly basis, hence it has to be transformed statistically to fit into our model.

The first data set is programmatically downloaded with an API call with a private key authenticating it. `wget` is used in downloading the data within Shell script. The later ones are downloaded manually from Google and Etherscan sites manually. The volume of the dataset is low, however, the volume increases as the consumption are initiated in real-time.

[Figure 3 about here.]

Figure 3 describes the snapshot layout of one of the source data. It has 8 columns about the Bitcoin statistics segregated on per day basis. The first column is the date at which the other columns are recorded, the second is the opening price of the Bitcoin compared to USD on that day, likewise third, fourth and fifth columns pertain to high, low and closing rates of Bitcoin. The sixth column represents BTC's transaction count on that day and seventh is the volume in terms of USD value, at last is the weighted price

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Figures 4, 5, 6 are the other features gathered from Google and Ethereum's mining pool.

## 5.3 Data Cleansing

The data cleansed with multiple Python and feature cleansing libraries in Python and Pyspark. Major efforts of cleansing are needed to standardize the date columns from all the data sources. The date format was in the different format in different sources. To stitch back all the data points, Date Time libraries were used and joined with a single standard format. Another important activity in the cleansing is data miss. For some instance, the values are missing resulting in incorrect predictions and correlations. To resolve these missing values, Imputer [11] functionality is used from feature library of Apache Spark. The imputer is an Imputation estimator for completing missing values, either using the mean or the median of the columns in which the missing values are located. The input to this function is dataframe columns and output are renamed dataframe columns. The processing happens in-memory with the spark.

## 5.4 Data Visualization

The visualization is provided in the form of static plots. Static plots are built-dimensional plots and scatter plots to represent correlation and projections.

## 6 SPEARMAN'S CORRELATION

Spearman's correlation function is used to identify the correlation between Bitcoin's price and the features selected. In Spark, a separate function is defined to calculate Spearman's correlation. The input is in Pyspark RDD's and the output value is returned between -1 to +1. The positive ratio indicates the feature is directly proportional and the negative values indicate indirect proportionality.

Spearman's Correlation on the selected features are :

- btc-vol :0.348540857386
- high :0.998581861669
- low :0.995190604708
- open :0.997943642437
- Google-trend:0.260343238604
- ETH price :0.68683414787
- ETHTRAN :0.720031468617
- btc-price-Label:1.0

[Figure 7 about here.]

The Figure7 describes the correlation between Bitcoin transaction count and its value. Or in other perspectives, the transaction count is consistent and due to the increase in price people started buying Bitcoins leading to volatility.

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

The Figures 8,9 and 10 describes the correlation between open, low and high prices of Bitcoin on the recorded date. This is obvious that the closing price is highly correlated with them.

[Figure 11 about here.]

As described in figure 11, the correlation plot provides the pattern between the search trends and the hike in price, which is clearly evident.

[Figure 12 about here.]

[Figure 13 about here.]

Figure 12 and 13 describes the correlation of Ethereum's price and its transaction volumes with Bitcoin's price in which the transaction pattern of Ethereum is more similar to Bitcoin's pattern is evident.

As far as processing is concerned, all the RDDs are cached before feeding into Spearman's correlation function, the reason being, when the RDDs are transformed multiple times, it has to calculate data lineage everytime it is computed and lineage is the basic quality of resilience in Apache Spark. If the RDDs are cached and persisted in-memory, the iteration and other transformations happen in memory avoiding costly I/O operations, this feature cannot be easily implemented when executed in conventional python libraries.

## 7 DECISION TREE REGRESSION

With the availability of features, we can take the processing to the next level of predicting Bitcoin's price. Here, supervised learning model is used to predict the price of Bitcoin.

[Figure 14 about here.]

Figure 14 give some basic idea of how decisions are made with the supervised decision tree based model.

Ensemble method models are derived from another base model. The base model used here is Decision trees and ensemble models are Random forest and Gradient Boosted tree(GBT) Algorithms.

Though the base model for both the algorithms is same, both are different in terms of training the dataset. GBT can train only one tree at a time whereas Random forest can train multiple trees resulting in reduced overfitting caused by GBTs.

"Random forests or random decision forests operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of mean prediction (regression) of the individual trees and GBTs iteratively train decision trees in order to minimize a loss function. Like decision trees, GBTs handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions" [12].

All the execution is implemented in Apache Spark, hence all the transformation and processing happens in-memory, even if the data volume is high, the processing will spawn across the clusters and will be processed with consistent redundancy.

The model is implemented by first splitting the data into two sets of different volume, i.e test data and training data. The training data will be used by the model to derive the logic and the built logic will be tested with the test data for accuracy. Here, 70:30 ratio is selected for training and test data respectively. And by altering this ratio we can adjust the performance of the model. Upon completion of the modeling, the accuracy of the models is calculated based on Metrics library in Spark. The metrics identified for the accuracy calculations are mean Squared Error, Root Mean Squared Error, r-square and mean Absolute error. Mean Squared error can be defined as an estimator to measures the average of the squares of the errors or deviations [13]. "R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression" [3].

## 7.1 Random Forest

As we are predicting Bitcoin's USD value per day, Bitcoin's price is considered as a label and all other columns are marked as features, and only the features having a decent level of correlation is marked as Features. These features are loaded into the model with Labelled Points as a Spark dataframe.

Random forest requires parameters to tune the model for the highest accuracy. Some of them are :

- Training dataset RDD of LabeledPoint
- NumTrees = 10 Number of trees in the random forest
- FeatureSubsetStrategy = auto Number of features to consider for splits at each node.
- Impurity = variance Criterion used for information gain calculation
- MaxDepth = 30 Maximum depth of tree
- MaxBins default Maximum number of bins used for splitting features
- Seed default Random seed for bootstrapping and choosing feature subsets

The first parameter is the RDD, which is the training data set, the next one is the number of trees, as the decision tree is based

on deriving mean of multiple decision trees, in general, the more trees gives better results. However, the improvement decreases as the number of trees increase more than the threshold of the given dataset. The FeatureSubsetStrategy defines how the features are sampled at each split in a tree. The Impurity parameter is the criteria followed for Information gain calculation, variance is the default considered by Spark. The next parameter is MaxDepth, which defines the limit of the depth of the tree, beyond which decision tree will not be extended. The next one is MaxBins, which describes the number of bins used for splitting and the last one is Seed parameter, which induces randomness while multiple trees are created.

With all parameters were carefully selected and the model is tuned to give the highest accuracy, Avg.closeness index of the algorithm is closer to 0.95. After deriving the model, the closeness/correctness of the predicted results was also analyzed and it is described in the plot 15.

## 7.2 Gradient Boosted Tree

In the Gradient Boosted algorithm, the training and test data are used in similar to the Random Forest algorithm. The implementation is less complex compared with Random forest. As GBT trains the model based on iterative execution of sequence of decision trees. Upon execution of three iterations, it is clearly evident with the closeness index that the data is little bit over-fitted with the closeness index of 0.96, slightly greater than Random forest.

The parameters used in these functions are

- Data Training dataset: RDD of LabeledPoint
- CategoricalFeaturesInfo Map storing arity of categorical features
- Loss Loss function used for minimization during gradient boosting
- NumIterations Number of iterations of boosting
- LearningRate Learning rate for shrinking the contribution of each estimator
- MaxDepth Maximum depth of tree
- MaxBins Maximum number of bins used for splitting features.

Important parameters used in GBT's functionality are Loss function, NumIterations and Learning rate. Number iteration is the number of times the tree is iterated to derive the result, the learning rate is optional, only used to alter the learning rate. The final one is the loss function, which defines the loss function used for minimization of loss during gradient boosting.

By altering these parameters, the performance of the model can be optimized.

## 8 RESULTS

From the observation of scatter plots of regression model Random forest 15 and GBT 16, it is evident that GBT's single tree iterative model has predicted the values with over-fitting. Some predicted values are consistent with some particular time scope and changes happening in steps. The prediction distribution looks like a single line and not widespread. Whereas, in the Random forest, the predicted values are widespread and closely aligned.

[Figure 15 about here.]

[Figure 16 about here.]

We have came up with a metric called *Closeness indicator*, which tells us mean ratio of test and predicted label. If it is less than 100, then the predicted value is less than the actual and if it is more than 100, then the predicted value is more than the actual value. For both algorithms, the closeness index is near 100%, hence both predictions achieved optimal results. Other important metric includes r-square values which above 95% in both the cases, hence our model fits with the expectation and the parameter selected for the algorithm holds good. Also, the other correlation graphs explain that Bitcoin transaction count doesn't impact the Bitcoin's values. even the exchange data doesn't impact much though the correlation looks close, there is a possibility that owners are not using the Bitcoin for any day to day transactions instead they are using as an asset like Gold, and by retaining it, the demand for the Bitcoin coin further increases. Also, the new-coins can only be generated through mining and the growth is controlled, coins in circulation keep reducing, increasing its cost. This analysis clearly proves that Bitcoin bought are saved in the wallets are not used in the regular transaction much. Most of the Bitcoin's are retained to earn the profit over its demand and its growth.

## 9 CHALLENGES FACED

Most of the challenges are with the data and casting to the required data types as the correlation and regression functions need data either in float or double data types. The other challenge faced is the data source availability, though the Bitcoin network is open to the public, gathering all the statistics data from all the mining pool available in BTC infrastructure is tedious. And, in the Bitcoin network, we do not know which user performs the transaction, we have no open option to classify the user and identify the feature inducing that transaction. Due to the void of these inducing factors, we may need to assume few features and start the analysis with the correlation. And handling all these constraints along with Big data specifics in mind adds up to the challenge, thanks to Apache Spark which handles the data lineage and persistence through RDDs.

## 10 PROJECT STRUCTURE

Three folders are created, the first one is for scripts which retain the actual code to be executed and two Korn shell scripts to install dependencies and to download the required source files. The second one is the data folder which retains the data required for the model and correlation algorithm. And, the extract folder is to persist the plot figures extracted out of the python script. The versioning and multiuser synchronization is supported by Git.

## 11 IMPROVEMENT OPPORTUNITIES

There are a lot of improvement opportunities to be implemented in the project. One of them includes fetching the data in near real-time directly from the Mining-pool instead of a third party data service, the second one would be increasing the granularity of data which would increase the performance and the Spark would make more sense with that level of granularity and volume. The other improvement opportunities include gathering more features like illegal market transaction data, mining exchange data, wallet exchange data, world's inconsistency data which will increase correlation

factors and result in the accurate prediction of models. Other visualization opportunity includes real-time presentation capabilities with Big data at the back end. Matplot API has minimal options for real-time reporting which can be upgraded. Other important improvement opportunity includes implementing the prediction logic with Neural network based models like Long Short-Term Memory(LSTM) as decision tree based models sometimes fail to adapt to the changes based on their past experience. These LSTM based model keep the memory of the previous experience and improve the learning upon training.

## 12 CONCLUSION

With all prowess of Big data and its technologies, Blockchain technologies are not only evolving, it also equips humans with the opportunity to make the world more transparent, ethical and a viable place to live. Even these technologies help us to conserve the earth by reducing resource consumption like paper, audit trials and power. As the technology has evolved so far, it is expected to understand its growth story in terms of its microscopic level to push it to the next level of improvement. Such microscopic level of qualities was missed in legacy methods and Big data comes to the rescue in identifying those qualities and nurture them.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] bitcoinnetwork.com. 2011. The Best Bitcoin Mining Pools For Making Money. (2011). <https://www.bitcoinnetwork.com/bitcoin-mining-pools/>
- [2] Etherscan.io. 2017. Ethereum Transaction Growth Chart. (2017). <https://etherscan.io/chart/tx>
- [3] Jim Frost. 2013. Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit. (May 2013). <http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
- [4] GDAX. 2017. Bitcoin Exchange – Ethereum Exchange – Litecoin Exchange – GDAX. (Dec 2017). <https://www.gdax.com/>
- [5] Michael Hendricks. 2011. 21 million cap. (2011). <https://bitcointalk.org/index.php?topic=3366.msg47522>
- [6] Alphabet INC. 2017. bitcoin - Explore - Google Trends. (2017). <https://trends.google.com/trends/explore?q=bitcoin>
- [7] Satoshi Nakamoto. 2008. *Bitcoin: A Peer-to-Peer Electronic Cash System*. Technical Report. [bitcoin.org](https://bitcoin.org). 9 pages. <https://bitcoin.org/bitcoin.pdf>
- [8] quandl. 2016. Search – Quandl. (2016). <https://www.quandl.com/search?query=>
- [9] AojaZhao saacMadan, ShauryaSaluja. 2016. Automated Bitcoin Trading via Machine Learning Algorithms. paper. (2016).
- [10] Spark. 2016. Overview - Spark 2.2.0 Documentation. (2016). <https://spark.apache.org/docs/latest/>
- [11] Apache Spark. 2016. pyspark.ml package fi? PySpark 2.2.0 documentation. (2016). <http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html>
- [12] Apache Spark. 2017. Ensembles - RDD-based API - Spark 2.2.0 Documentation. (2017). <https://spark.apache.org/docs/2.2.0/mllib-ensembles.html>
- [13] Wikipedia. 2017. Mean squared error - Wikipedia. (2017). [https://en.wikipedia.org/wiki/Mean\\_squared\\_error](https://en.wikipedia.org/wiki/Mean_squared_error)
- [14] Wikipedia. 2017. Spearman's rank correlation coefficient - Wikipedia. (2017). [https://en.wikipedia.org/wiki/Spearman%27s\\_rank\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient)
- [15] WolfeZhao. 2017. *OSTK to HODL: Overstock to Keep 50% of All Bitcoin Payments as Investments - CoinDesk*. Technical Report. coindesk. <https://www.coindesk.com/ostk-hodl-overstock-keep-50-bitcoin-payments-investments/>
- [16] Xoom.inc. 2017. XE Money Transfer Tips: Why Do Currencies Fluctuate. (2017). <http://www.xe.com/moneytransfertips/why-do-currencies-fluctuate.php>

## LIST OF FIGURES

1	Spark Architecture	7
2	Project Architecture on Spark	8
3	Bitcoin mining statistics data	8
4	Google's trend data	9
5	Ethereum's pricing on daily basis	9
6	Ethereum transactions on daily basis	9
7	BTC Transaction and USD value - 0.348540857386	10
8	Bitcoin Highest exchange value and Closing value - 0.998581861669	10
9	Bitcoin Lowest exchange value and Closing value - 0.995190604708	11
10	Bitcoin Opening exchange value and Closing value - 0.997943642437	11
11	Bitcoin USD value and Google search trend - 0.260343238604	12
12	ETH price and BTC price - 0.68683414787	12
13	ETH Transaction volume and BTC transaction volume - 0.720031468617	13
14	Sample Decision tree	13
15	Randomforest Scatterplot	14
16	GBT Scatterplot	15

# Spark Architecture



Figure 1: Spark Architecture

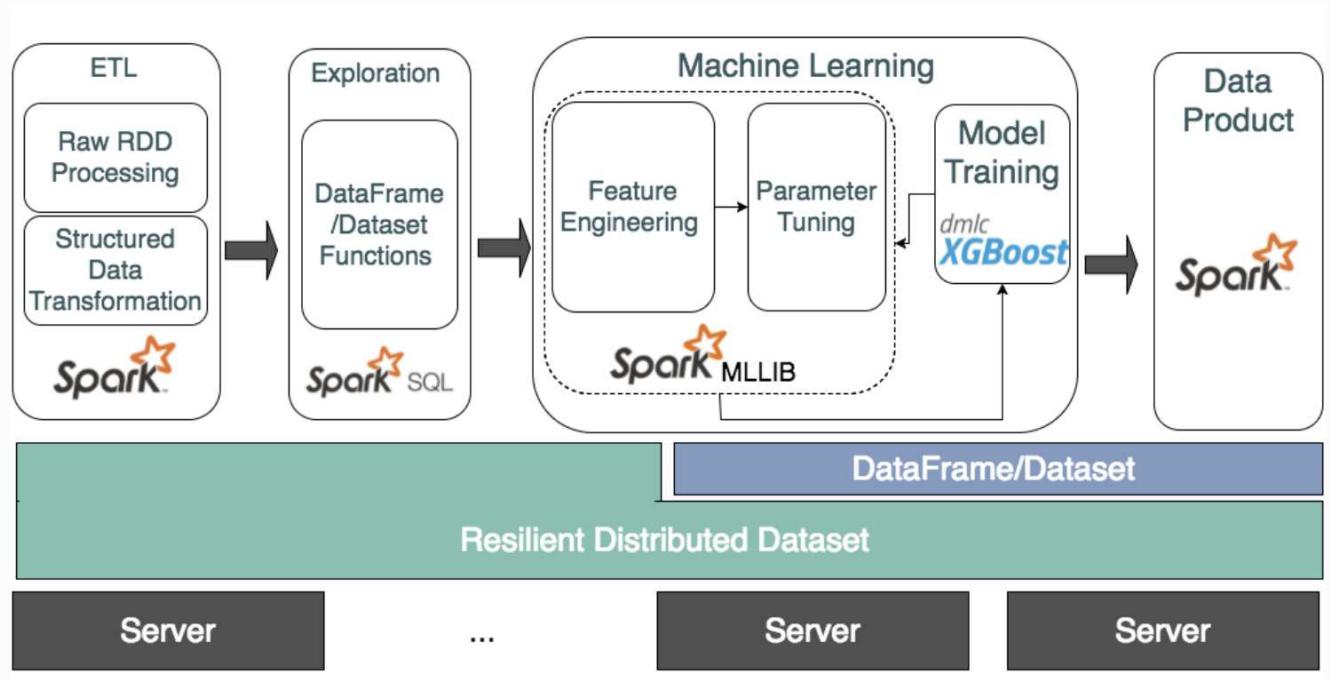


Figure 2: Project Architecture on Spark

```
Date,Open,High,Low,Close,Volume (BTC),Volume (Currency),Weighted Price
2017-11-29,9949.0,9949.0,9945.96,9945.97,20.29785801,201905.011261,9947.10925468
2017-11-28,9768.71,9989.95,9705.99,9949.0,18393.0818191,181959821.389,9892.84031783
2017-11-27,9401.11,9795.0,9401.01,9768.71,24642.0982679,237829776.608,9651.36061153
2017-11-26,8795.5,9596.0,8795.5,9401.11,27568.0196716,253816538.997,9206.91954012
2017-11-25,8215.01,8795.5,8203.98,8795.5,16239.7909316,138404574.912,8522.55891071
2017-11-24,8031.16,8324.0,7900.0,8215.01,14213.6286885,116296487.968,8182.04066794
2017-11-23,8250.0,8274.98,8031.16,8031.16,11685.6896442,95602227.3138,8181.13694823
2017-11-22,8109.0,8298.98,8103.13,8250.0,13107.9400888,107439055.55,8196.48661974
2017-11-21,8256.01,8375.0,7802.99,8109.0,29504.8066982,239617882.069,8121.31679154
2017-11-20,8031.83,8293.25,7969.0,8256.01,15479.8961619,126479984.708,8170.59645525
2017-11-19,7777.01,8098.62,7700.0,8031.82,14085.4833982,111595386.42,7922.72322253
2017-11-18,7714.7,7847.98,7502.0,7777.01,14531.7006631,111676672.782,7685.03806756
2017-11-17,7838.54,7988.5,7536.0,7714.71,23950.8477411,187290409.41,7819.78205677
2017-11-16,7294.0,7985.37,7130.0,7838.53,28404.999214,214993745.259,7568.8699598
2017-11-15,6605.0,7349.0,6605.0,7294.0,27327.1284653,193024708.734,7063.48305052
2017-11-14,6535.87,6748.0,6464.64,6605.0,19505.257774,128729345.115,6599.72539747
2017-11-13,5886.35,6841.45,5850.0,6535.87,35150.8905255,224596217.133,6389.4886808
2017-11-12,6246.64,6406.0,5511.11,5806.25,10610.2257667,200525162.810,6017.41077071
```

Figure 3: Bitcoin mining statistics data

```

2012-12-02,1
2012-12-09,1
2012-12-16,1
2012-12-23,1
2012-12-30,1
2013-01-06,1
2013-01-13,1
2013-01-20,1
2013-01-27,1
2013-02-03,1
2013-02-10,2
2013-02-17,2
---- -- -- -

```

Figure 4: Google's trend data

```

Date(UTC),UnixTimeStamp,Value
7/30/2015,1438214400,0.00
7/31/2015,1438300800,0.00
8/1/2015,1438387200,0.00
8/2/2015,1438473600,0.00
8/3/2015,1438560000,0.00
8/4/2015,1438646400,0.00
8/5/2015,1438732800,0.00
8/6/2015,1438819200,0.00
8/7/2015,1438905600,3.00
8/8/2015,1438992000,1.20
8/9/2015,1439078400,1.20
8/10/2015,1439164800,0.00
8/11/2015,1439251200,0.99
8/12/2015,1439337600,1.29
8/13/2015,1439424000,1.88

```

Figure 5: Ethereum's pricing on daily basis

```

Date(UTC),UnixTimeStamp,Value
7/30/2015,1438214400,8893
7/31/2015,1438300800,0
8/1/2015,1438387200,0
8/2/2015,1438473600,0
8/3/2015,1438560000,0
8/4/2015,1438646400,0
8/5/2015,1438732800,0
8/6/2015,1438819200,0
8/7/2015,1438905600,2050
8/8/2015,1438992000,2881
8/9/2015,1439078400,1329
8/10/2015,1439164800,2037
8/11/2015,1439251200,4963
8/12/2015,1439337600,2036
8/13/2015,1439424000,2842

```

Figure 6: Ethereum transactions on daily basis

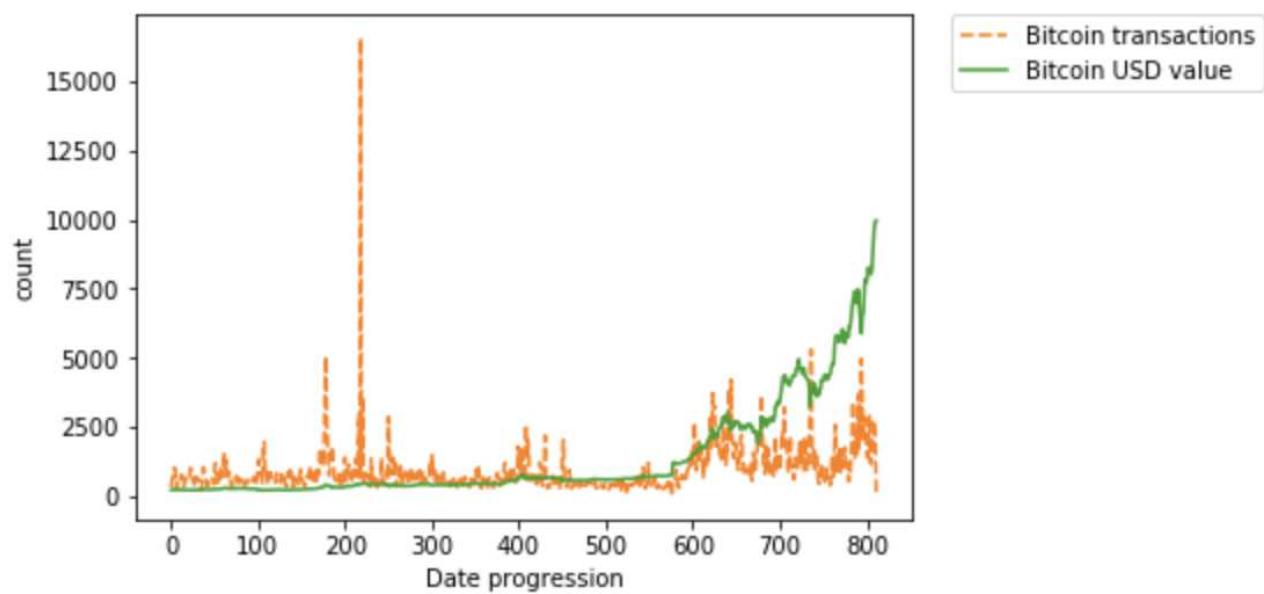


Figure 7: BTC Transaction and USD value - 0.348540857386

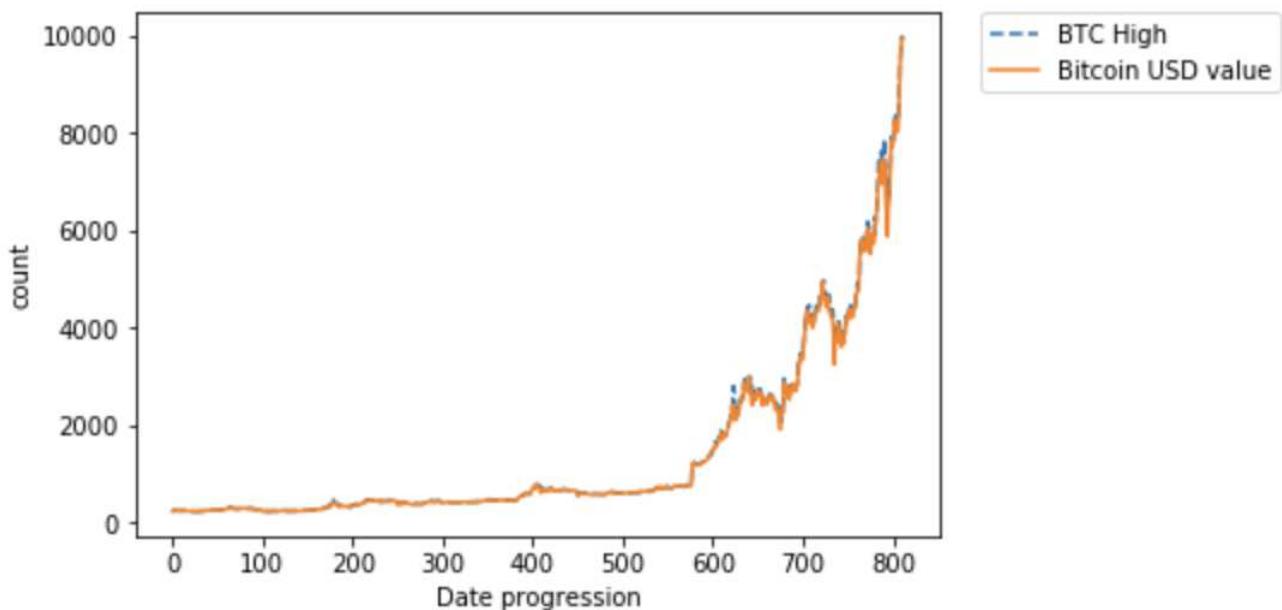


Figure 8: Bitcoin Highest exchange value and Closing value - 0.998581861669

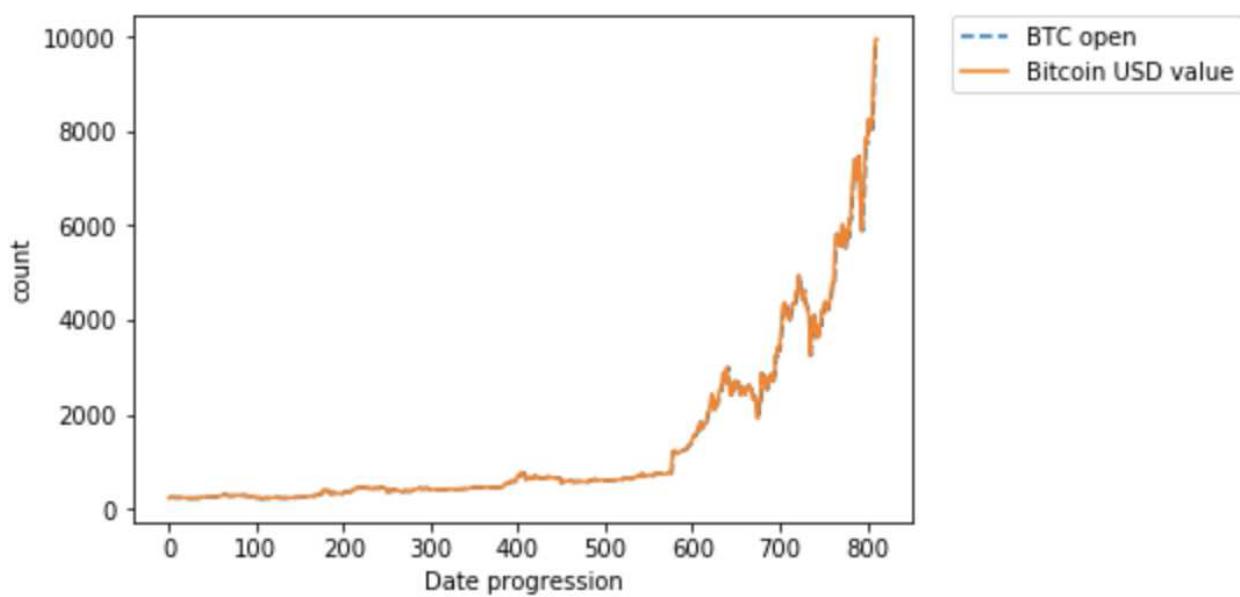


Figure 9: Bitcoin Lowest exchange value and Closing value - 0.995190604708

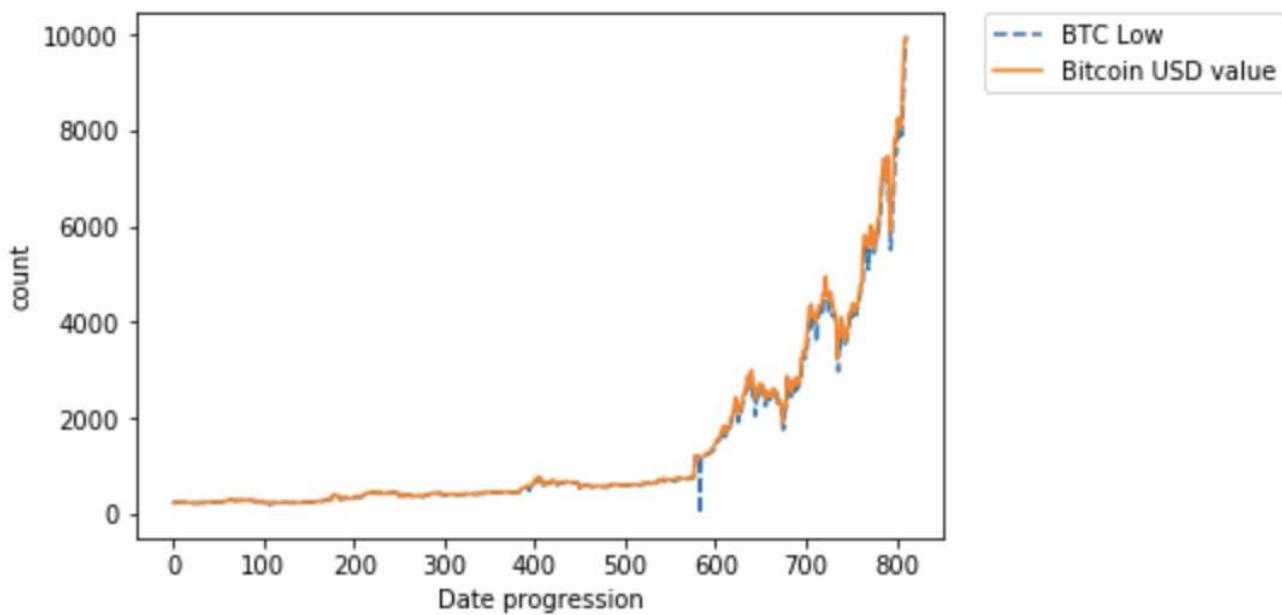


Figure 10: Bitcoin Opening exchange value and Closing value - 0.997943642437

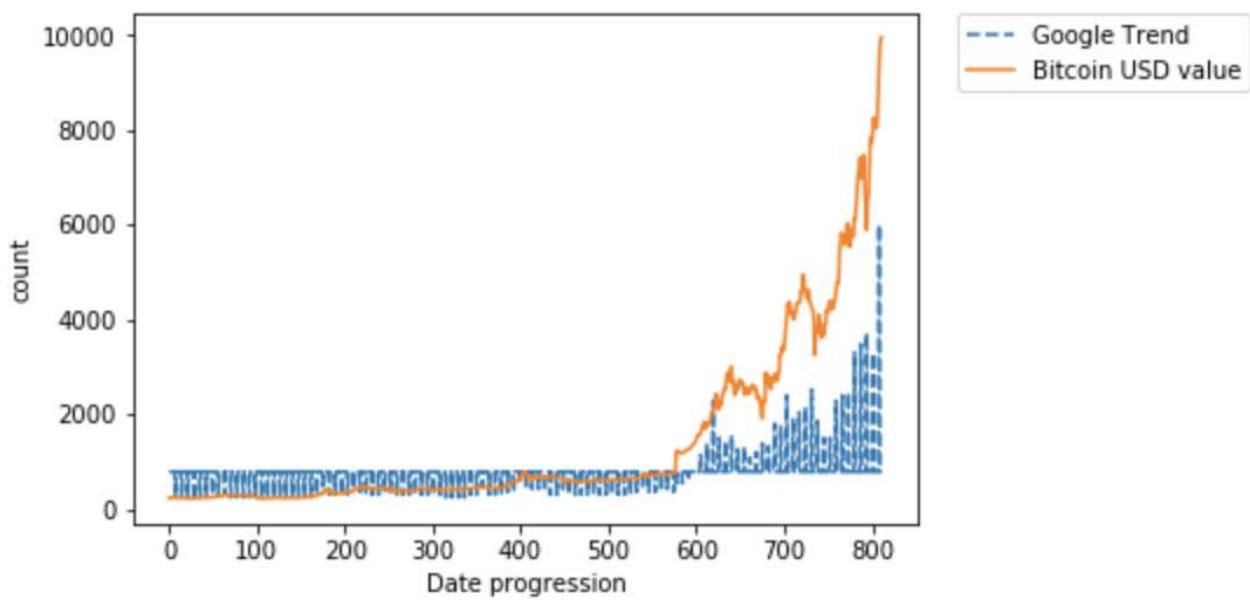


Figure 11: Bitcoin USD value and Google search trend - 0.260343238604

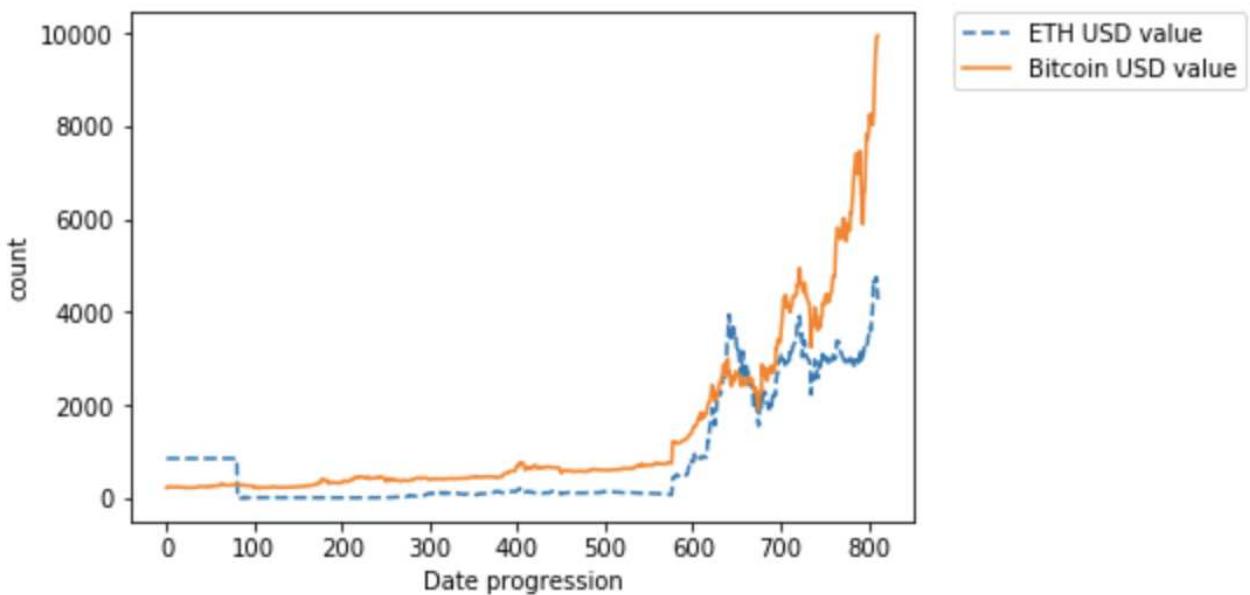


Figure 12: ETH price and BTC price - 0.68683414787

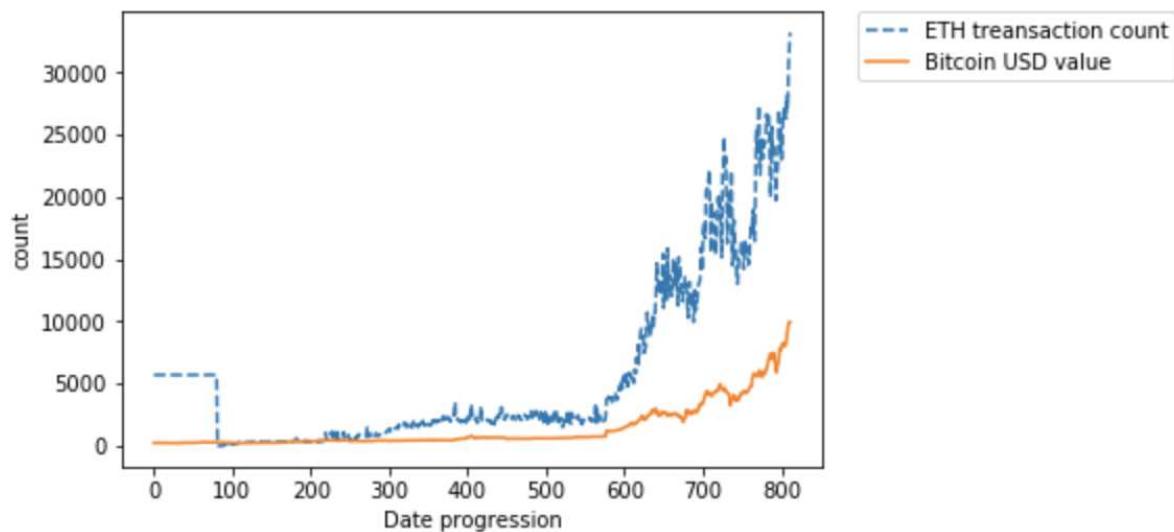


Figure 13: ETH Transaction volume and BTC transaction volume - 0.720031468617

`TreeEnsembleModel classifier with 3 trees`

```

Tree 0:
  Predict: 1.0
Tree 1:
  If (feature 0 <= 1.0)
    Predict: 0.0
  Else (feature 0 > 1.0)
    Predict: 1.0
Tree 2:
  If (feature 0 <= 1.0)
    Predict: 0.0
  Else (feature 0 > 1.0)
    Predict: 1.0

```

Figure 14: Sample Decision tree

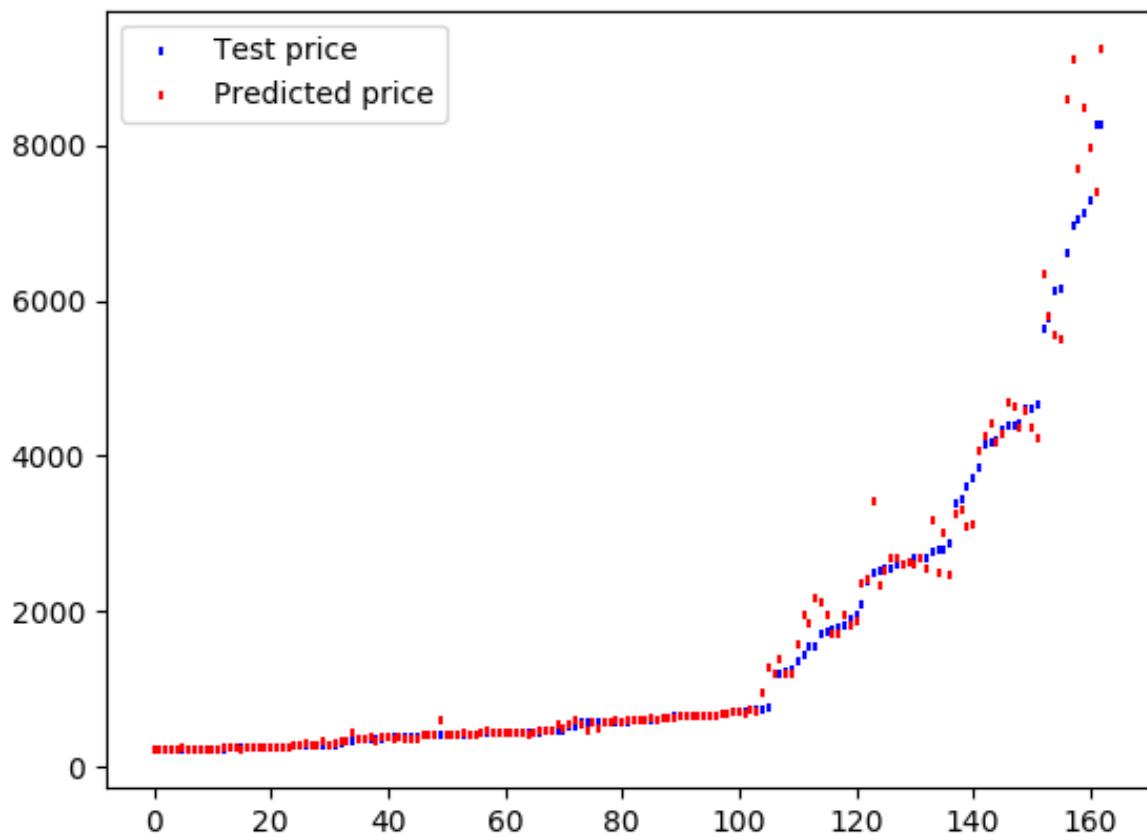


Figure 15: Randomforest Scatterplot

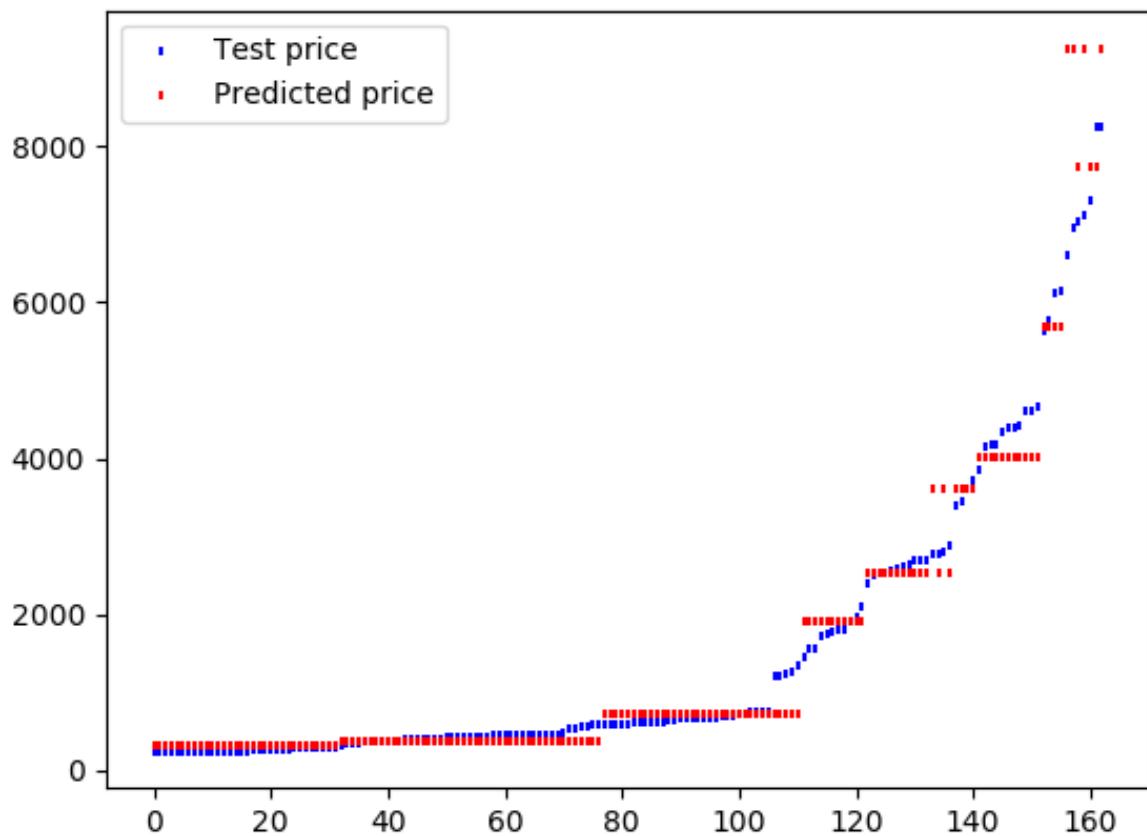


Figure 16: GBT Scatterplot

```
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:93: AUTHOR1_EMAIL = {satoshin@gmx.com},
```

```
bibtext space label error
```

```
=====
report.bib:93: AUTHOR1_EMAIL = {satoshin@gmx.com},
```

```
bibtext comma label error
```

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

```
[2017-12-10 13.51.35] 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
8:81      error    line too long (142 > 80 characters) (line-length)
20:81     error    line too long (117 > 80 characters) (line-length)
38:81     error    line too long (388 > 80 characters) (line-length)
```

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

```
name: Ashok Kuppuraj
```

```
hid: 324
```

```
paper1: 100% Oct 31 17
```

```
paper2: 100% Nov 6 17  
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)  
wc 324 project (null) 4538 report.tex  
wc 324 project (null) 4683 report.pdf  
wc 324 project (null) 392 report.bib
```

```
find "
```

---

60: Feature extraction, transformation, and prediction can be synonymous with a conventional ETL methodology. Though few of the extraction is handled manually and the volume is comparably low, it is assumed that the data volume will be increased by modifying the extraction to real-time systems. When the extraction systems are changed, our code must be able to handle streaming data which can be related to "variety and volume " of the data. The next step is validating the data for anomalies, data miss and cleanse the data of issues which is synonymous with data cleansing. The later one is data processing, which includes data processing with multiple iterations and permutation consuming a lot of memory and other resources. These processing needs lead us in adopting Big data technologies in the entire lifecycle of the implementation.

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

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

---

```
passed: False
```

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

---

```
passed: False

floats
-----

66: \begin{figure}![ht]
67: \centering\includegraphics[width=\columnwidth]{images/Sparkarchic.
  png}
69: \label{fig:sparkarchi}
72: Spark's architectures given in the figure \ref{fig:sparkarchi}
  provide a glimpse of how different system in Spark is interfaced.
  The first level of interfacing to Spark is with high-level
  languages like Scala, Java, Python and R. Users implement their
  functionalities in these high -level languages. The primary
  executing components in Spark are Driver and Executor modules. The
  driver is the entry point for any implementation, the written
  programs will be executed in the main function of the Driver
  module, later converted to set of Directed Acyclic graph by the
  Spark APIs. DAGs are then executed in executors in the data nodes
  based on the data placement policy of the infrastructure. Four
  modules built on spark for serving the user's needs are SparkSQL,
  Spark Streaming, MLLib, and GraphX. Spark SQL and Machine Learning
  libraries(MLLib) are consumed in our implementation and the future
  improvement would be on Spark Streaming which is used for
  Streaming requirements. SparkSQL and MLLibs modules contain the
  implementation for DataFrames, SQL functionalities, and Machine
  learning libraries. The next level of the modules is data
  abstraction layer. Spark's basic data abstraction is Resilient
  Distributed Dataset (RDD), which is a fault tolerant partitioned
  data encapsulation datatype. The RDDs are lazily evaluated, hence
  a Directed Acyclic Graph is implemented to persist the state of
  the RDDs at each stage. With RDD, Spark can execute the
  transformation in parallel with fault tolerance. This
  implementation widely differentiates from conventional Python
  implementation which lacks this advanced logic.
78: The architecture flow consists of three levels of components,
  first one is the Data extraction, second is processing and the
  final is visualization. The Figure \ref{Architecture:project}
  describes how the implementation is fitted over Spark's
  architecture.
81: \begin{figure}![ht]
82: \centering\includegraphics[width=\columnwidth]{images/Projectflow.
  png}
84: \label{Architecture:project}
107: \begin{figure}![ht]
```

```

108: \centering\includegraphics[width=\columnwidth]{images/Source1data
    .png}
110: \label{fig:sourcedata}
114: Figure \ref{fig:sourcedata} describes the snapshot layout of one
    of the source data. It has 8 columns about the Bitcoin statistics
    segregated on per day basis. The first column is the date at
    which the other columns are recorded, the second is the opening
    price of the Bitcoin compared to USD on that day, likewise third,
    fourth and fifth columns pertains to high, low and closing rates
    of Bitcoin. The sixth column represents BTC's transaction count
    on that day and seventh is the volume in terms of USD value, at
    last is the weighted price
116: \begin{figure}[!ht]
117: \centering\includegraphics[width=0.18\columnwidth]{images/googled
    ata.png}
119: \label{fig:googledata}
122: \begin{figure}[!ht]
123: \centering\includegraphics[width=0.4\columnwidth]{images/ethprice
    .png}
125: \label{fig:ethpri}
128: \begin{figure}[!ht]
129: \centering\includegraphics[width=0.4\columnwidth]{images/ethtran.
    png}
131: \label{fig:ethtran}
134: Figures \ref{fig:googledata}, \ref{fig:ethpri}, \ref{fig:ethtran}
    are the other features gathered from Google and Ethereum's mining
    pool.
158: \begin{figure}[!ht]
159: \centering\includegraphics[width=\columnwidth]{images/BTC-
    prcvsBTC-trans.png}
161: \label{1}
164: The Figure\ref{1} describes the correlation between Bitcoin
    transaction count and its value. Or in other perspectives, the
    transaction count is consistent and due to the increase in price
    people started buying Bitcoins leading to volatility.
166: \begin{figure}[!ht]
167: \centering\includegraphics[width=\columnwidth]{images/High.png}
169: \label{2}
172: \begin{figure}[!ht]
173: \centering\includegraphics[width=\columnwidth]{images/Open.png}
175: \label{3}
178: \begin{figure}[!ht]
179: \centering\includegraphics[width=\columnwidth]{images/low.png}
181: \label{4}
184: The Figures \ref{2},\ref{3} and \ref{4} describes the correlation
    between open, low and high prices of Bitcoin on the recorded

```

date. This is obvious that the closing price is highly correlated with them.  
 186: \begin{figure} [!ht]  
 187: \centering\includegraphics[width=\columnwidth]{images/googletrend.png}  
 189: \label{5}  
 192: As described in figure \ref{5}, the correlation plot provides the pattern between the search trends and the hike in price, which is clearly evident.  
 194: \begin{figure} [!ht]  
 195: \centering\includegraphics[width=\columnwidth]{images/ethvalue.png}  
 197: \label{6}  
 200: \begin{figure} [!ht]  
 201: \centering\includegraphics[width=\columnwidth]{images/ethtrancount.png}  
 203: \label{7}  
 206: Figure \ref{6} and \ref{7} describes the correlation of Ethereum's price and its transaction volumes with Bitcoin's price in which the transaction pattern of Ethereum is more similar to Bitcoin's pattern is evident.  
 214: \begin{figure} [!ht]  
 215: \centering\includegraphics[width=0.75\columnwidth]{images/Decisiontree.png}  
 217: \label{fig:8decisiongree}  
 221: Figure \ref{fig:8decisiongree} give some basic idea of how decisions are made with the supervised decision tree based model.  
 252: With all parameters were carefully selected and the model is tuned to give the highest accuracy, Avg.closeness index of the algorithm is closer to 0.95. After deriving the model, the closeness/correctness of the predicted results was also analyzed and it is described in the plot \ref{scpl:ran}.  
 274: From the observation of scatter plots of regression model Random forest \ref{scpl:ran} and GBT \ref{scpl:gbt}, it is evident that GBT's single tree iterative model has predicted the values with over-fitting. Some predicted values are consistent with some particular time scope and changes happening in steps. The prediction distribution looks like a single line and not widespread. Whereas, in the Random forest, the predicted values are widespread and closely aligned.  
 276: \begin{figure} [!ht]  
 277: \centering\includegraphics[width=\columnwidth]{images/RandomForestsscatterplot.png}  
 279: \label{scpl:ran}  
 282: \begin{figure} [!ht]  
 283: \centering\includegraphics[width=\columnwidth]{images/GBTscatterplot}

```
    lot.png}
285: \label{scpl:gbt}

figures 16
tables 0
includegraphics 16
labels 16
refs 11
floats 16

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 above may be used improperly

289: For both algorithms, the closeness index is near 100\%, hence  
both predictions achieved optimal results. Other important metric  
includes r-square values which are above 95\% in both the cases,  
hence our model fits with the expectation and the parameter  
selected for the algorithm holds good.

---

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

# Predicting Profitable Customers in Banking Industry

Dhanya Mathew  
Indiana University  
711 N Park Ave  
Bloomington, Indiana 47408  
dhmathew@iu.edu

## ABSTRACT

Banks often want to know the profile of their profitable top 1% or 20% customers looks like. Conversely, they may also wonder what the general profile is of the customers in the worst 1% and 20% of profit. Based on customer's data variables at any given time, a good predictive model can predict which profit group (extremely unprofitable, average, extremely profitable etc) customers fall into. This helps financial institutions to better understand what drives the customer profit and accordingly take decisions to sell their products to the right customers. Further down in banking sector, it is a challenge to identify customers who are most likely to repay the loan. Recent big data and machine learning technologies have the potential to predict good customers and open doors for banks to profitable growth. Since the banking sector has evolved over the periods, there are tremendous amount of historical data available to analyze. We show how bank's big data can be analyzed and create a model based on that, to classify customers. In addition to big data technologies, we use machine learning algorithms to build a predictive model to predict creditworthy and uncreditworthy customers from a list of new customers. Various classification algorithms like Decision Tree and Random Forest are used to build the models and trace the best model among them to achieve the goal.

## KEYWORDS

i523, HID328, Big Data, Spark, Python, Decision Trees, Random Forest

## 1 INTRODUCTION

Big data as the name implies, refers to large and complex data which continues to grow enormously day by day. Industries like financial firms, in particular, have widely adopted big data analytics to obtain better investment decisions with consistent growth. Recent survey research indicates that 71 percent of firms in the financial services industry at a global level are exploring big data and predictive analytics [22]. This number continues to grow and sectors like government, business, technology, universities, health-care, finance, manufacturing etc make use of big data to obtain meaningful information using big data technologies [31].

The finance sector contributes to the daily data generation from products and marketing, banking, business, share market etc [14]. Banking is a very sensitive field and any useful insight can make a positive impact on the overall turnover. Historic data analysis and real time data analysis are equally important in banking sector. The era of big data helps financial firms to take quality business decisions related to expanding revenues, managing costs, hiring resources etc, based on effective data analysis which provide access

to real-time insights. Data-driven decision making is one of the key advantages of big data technologies.

### 1.1 Project Goals

We aim to help banking sector to identify trustworthy customers. Specifically, help banks to take a decision driven by data, whether to approve or reject a loan application. When a new customer approaches the bank for a loan, banks would be able to identify the customers who are most likely to repay the loan by analyzing the applicant's profile and background information.

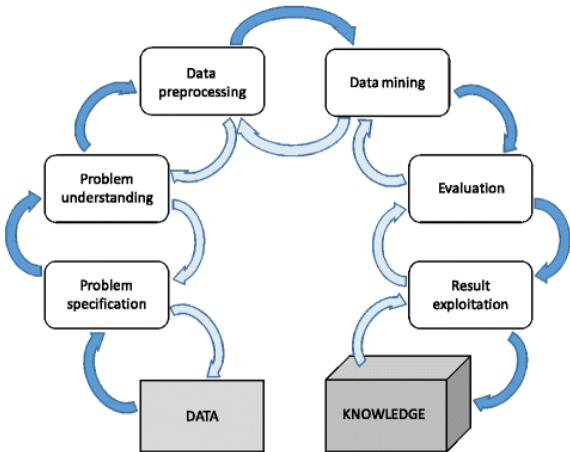
There can be two scenarios of risks associated with the bank's decision. First, if the customer is creditworthy and if the bank rejects the loan, then it is a loss to the bank in terms of interest. Second, if the customer is uncreditworthy and if the bank approves the loan, then it is a loss to the bank in terms of loan amount and interest [15]. Approving loan for an uncreditworthy customer will end up in more financial loss for the bank and accordingly is a greater risk. Hence banks would require a decision rule to follow for whom to approve the loan. We show how to build a predictive model using machine learning algorithm and a sample dataset with customer records classified as "Good" or "Bad" according to bank's opinion on the customer. With our model, we try to mitigate these risks for the banks by contributing to the decision rules. In other words, our model helps to minimize the risks and maximize the profit for the bank by understanding the customers.

### 1.2 Methods and Technologies Involved

The goal of most of the big data projects is to analyze the data and derive knowledge out of it. In other words, data is the input to the model and knowledge is the output. We also follow the same methods and processes for our project. We wrote the project code using Python3.

*1.2.1 Project Workflow.* Overall workflow of the project is shown in Figure 1. We have taken a sample data set of loan applications received by a bank. We explored the data and the requirements of the bank and based on that set the project goals as discussed in the section 1.1, before starting the project. In the real scenarios, we will not be able to apply analytical methods directly on the raw data as it likely be imperfect and containing irrelevant information. Hence we do data cleaning (data preprocessing) as the first step. Data cleaning is done using PySpark. The cleaned data has 1000 customer records with 1 classifier and 20 feature variables.

Exploratory analysis like Chi-square test is done to understand the data and feature selection for analysis as part of the data mining process. We have done graphical representations to show how the actual data is related and what are the direct insights available from the cleaned data set.



**Figure 1: Project Workflow [9]**

Machine learning approaches are used for the data evaluation and to build our predictive model. We develop various models using machine learning algorithms and compare them to identify the best model to choose for our problem solution [23]. To develop the models we first split the data set into two parts - training data and test data. We defined 2 baseline models, Decision trees and the Random Forest model. We compare all these models to identify the most effective and least penalty model. We use python as the programming language to build these models and display visualizations for easy comparison and results discussion.

**1.2.2 Python.** Python version 3 is the programming language used to develop the models and visualizations in this project. Python is a general purpose programming language that is open source, easy to use, faster to write, flexible and powerful. It has a rich set of libraries and utilities for data processing and analytics tasks [27]. Other important features of Python include the ability to process big data, scalability of applications and easiness to integrate with web applications. We use Python libraries like pandas, matplotlib, seaborn and numpy.

**Pandas:** Pandas is one of the most popular libraries in Python. Pandas is used for data manipulation and analysis [17], read data files from different sources, create data frames and some built-in visualizations [11].

**Matplotlib:** Matplotlib is the library used for plotting arrays and histograms of data in python [6].

**Seaborn:** Seaborn is a Python visualization library used for statistical visualization of data [30].

**Numpy:** Numpy is Python library which is used to operate mathematical functions on large multi dimensional arrays [34].

**1.2.3 PySpark.** Even though Python is powerful to handle complex big data analytic tasks, it alone cannot handle the big data processing. A distributed framework would require to handle a large amount of data. Spark is a distributed computing framework which supports Python [7].

PySpark is used to carry out the data preprocessing tasks and it is the Python API for Spark.

**1.2.4 Jupyter Notebook.** Jupyter notebook is an open source web application that allows to edit, run and share Python code and visualizations into a web view. It can be used to modify and re-execute program parts in a flexible way [5]. The files created in Jupyter notebook use extension ".ipynb".

**1.2.5 Machine Learning.** Machine learning enables computers to learn automatically and act accordingly without human assistance or being explicitly programmed. It is an application of Artificial Intelligence. It focuses on computer programs that can access data and learn by itself. Learning process starts by observing the data for patterns and make better decisions in future on the given scenarios [25]. There are mainly 2 categories of machine learning - Supervised and Unsupervised.

**Supervised Machine Learning Algorithms:** Supervised machine learning algorithms enable machines to get trained using a known training data set. Using these labeled examples, supervised learning algorithms can predict future events by applying already learned knowledge. These systems can be used for target definitions for new set of data after required training. Also, it can compare new data input with the intended output and give error indications [25].

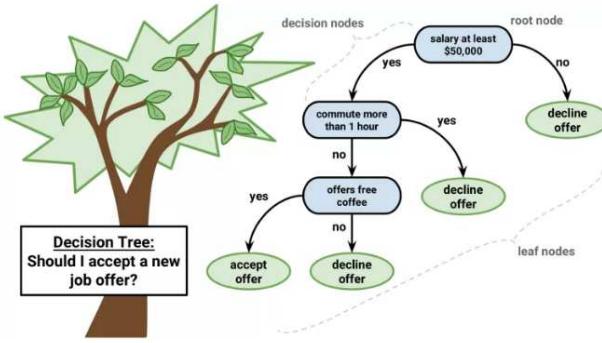
**Unsupervised Machine Learning Algorithms:** Unsupervised machine learning algorithms are used when there is not a preferred output and the data is not labeled or classified. It helps to find the hidden patterns in the data. It can describe the hidden structure of the unlabeled data but would not be useful to provide a correct, intended output [25].

There is another categorization of the machine learning algorithms depending on the preferred output. That include *Classification Algorithms* (Used for supervised learning with discrete output), *Regression Algorithms* (Used for supervised learning with continuous output), *Clustering Algorithms* (Unsupervised) etc [33].

We use supervised machine learning approaches in this project. In particular, the classification algorithms - Decision Tree and Random Forest.

**1.2.6 Decision Tree.** Decision Tree is a supervised machine learning algorithm used to solve both classification and regression problems. In Decision Tree, a trained model with a set of rules will be created based on the training data. The target class or value of a test/new data set will be predicted based on this training rules. Decision Tree algorithm is simple to understand as it uses a tree model representation to solve the problem. It starts from a root node and continues with other decision nodes. Each internal decision nodes corresponds to the feature variables and each leaf nodes corresponds to the class label [24]. Figure 2 shows the decision tree classifier.

The best attribute will be chosen as the root node. To identify the root node there are 2 methods. They are, *Information Gain* and *Gini Index*. There are statistical approaches to calculate Information gain and Gini index values for each feature variables. Attribute with better value will be considered as the root node and other attributes will be placed in the internal nodes according to the values in recursive order. Step 1 to model the decision tree is placing the root attribute. In step 2, the training data set will be divided into

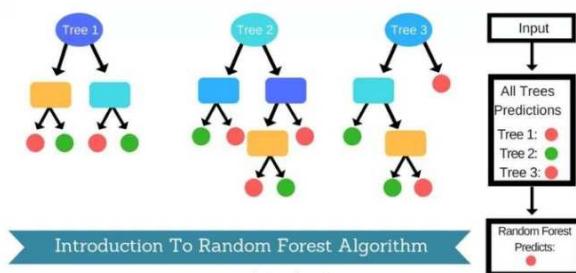


**Figure 2: Decision Tree Classifier [24]**

2 sub data sets in such a way that, both subsets will contain same attribute values for that variable. Step 1 and 2 will be repeated until we reach the leaf nodes with predicted class value [24].

*Overfitting:* Overfitting is a practical issue that can happen while building a decision tree. When the algorithm goes deeper and deeper it builds more branches because of the irregularities in data and the prediction accuracy of the model goes down accordingly. There are 2 methods can be used to avoid overfitting issues - *Pre-Pruning* and *Post-Pruning*. In Pre-pruning, we set a threshold value as a goodness measure and if it crosses, further split of the node will be stopped. In Post-Pruning, tree construction continues until all leafs are reached and pruning will be done if the model shows overfitting issues. Cross-validation data will be used to measure the improvement in this method [24].

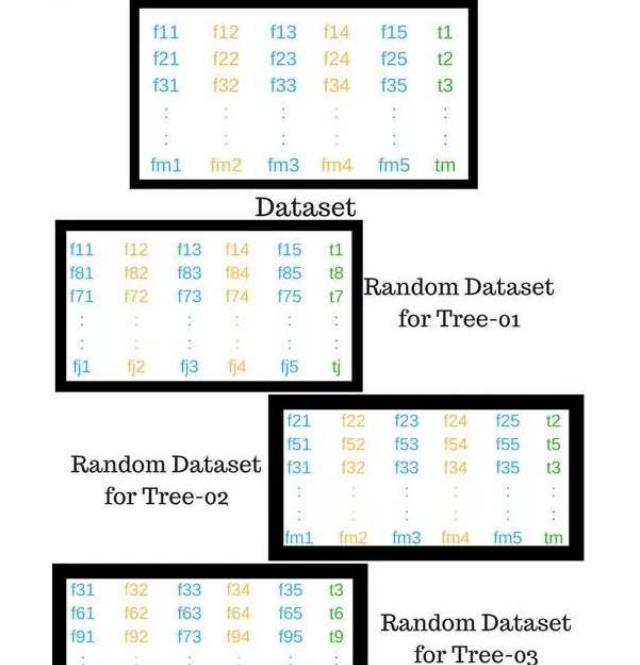
**1.2.7 Random Forest.** Like Decision Tree, Random Forest algorithm also can be used for classification as well as regression problems. It is a supervised machine learning algorithm. It uses decision tree concept as well but there will be more than one trees in a Random Forest. As the number of trees increases in the Random forest, the accuracy of the prediction also will increase accordingly. Random forest algorithm can handle missing values in the data. Also with more trees in the forest, overfitting issues will not occur in Random Forest algorithm [19]. Figure 3 shows the Random Forest model.



**Figure 3: Random Forest model [19]**

Random Forest algorithm progresses via 2 stages - Random Forest creation and Perform prediction. To create the Random Forest, we select a random number of feature variables from the total list of

feature variables in the training data and create a Decision Tree out of it. We repeat this process to create desired number of trees. These randomly created trees will form a Random Forest. Figure 4 shows how random forest algorithm works.



**Figure 4: How random forest algorithm works [19]**

The test data set will be analyzed against the rules developed by each of the trees to predict the output. To predict Random Forest output, the outputs of each of the trees are considered as votes. The top voted output is the final predicted value of the Random Forest.

### 1.3 Installations

Technologies used in this project are discussed in detail in section 1.2. The installation commands on Ubuntu 16.04 OS for each of these technologies are given in this section. Installations can be done from the terminal window.

- (1) Python installation steps for ubuntu OS are available in the askubuntu website [4].
- (2) Install pip to manage the libraries in the Python. Pip is a Python package management software used to install and manage Python libraries. pip can be installed using command "sudo pip install -U pip" [18].
- (3) Install PySpark using command "sudo pip install pyspark" [1].
- (4) Install Jupyter notebook using command "sudo pip install jupyter" [20].
- (5) Install Pandas using command "sudo pip install pandas" [16].
- (6) Install matplotlib using command "sudo pip install matplotlib" [3].
- (7) Install seaborn using command "sudo pip install seaborn" [26].

**Table 1: Variables and description [15]**

Variable	Description
Credit	Creditability: Good or Bad
Account Status	Balance of current account
Credit Months	Duration of Credit (month)
Credit History	Payment Status of Previous Credit
Purpose	Purpose of credit
Credit Amount	Amount of credit
Savings	Value Savings or stocks
Employment	Length of current employment
Installment Rate	Installment in % of current income
Personal Status	Sex and Marital Status
Guarantors	Further debtors
Residence	Duration in Current address
Property	Most valuable available asset
Age	Age in years
Other Installments	Concurrent Credits
Housing	Type of apartment
Credit Cards	No of Credits at this Bank
Occupation	Occupation
Dependents	No of dependents
Telephone	Phone number
Foreign Worker	Foreign worker

(8) Install numpy using command "sudo pip install numpy" [2]

All these installation steps are included in a make file referred in appendix A.1.

## 2 DATASET

We used the German Credit data which is publicly available in the UCI Machine Learning Repository [10] and also in the website of PennState Eberly College of Science [15]. Both these sites have the cleaned dataset and not the original one. The dataset that we used (german-credit.csv) is taken from the website of PennState Eberly College of Science [15] and it is uploaded in the Github repository [12]. We recreated the original one from these data sets to understand and try out the data cleaning processes. We start our project with the recreated original data set (credit-data.csv) which is available in the Github repository [12].

Dataset includes 1000 customer records with 20 feature variables and a class variable. In the class variable, the actual class of the customer is specified - good or bad. The complete list of data set variables and their description is given in Table 1.

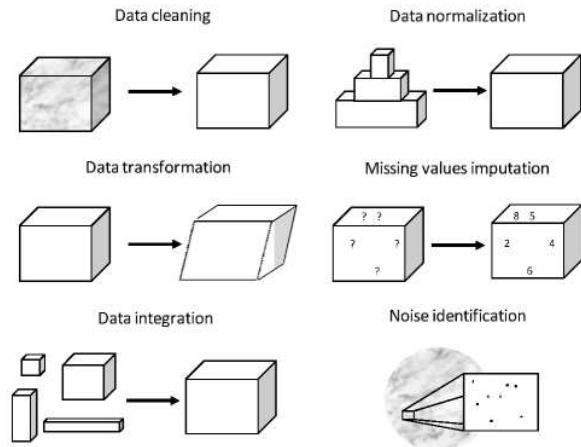
## 3 DATA CLEANSING

A massive amount of raw data is piling up in the recent years from different sources and it has been continuously getting stored as the storage mechanism is getting cheaper and the storage capacity increases day by day. This raw data cannot be analysed as it is by human or traditional applications, as the processing capacity of traditional tools has been exceeded because of the volume of the data. That is the reason why big data technologies have evolved and they use distributed systems like MapReduce, Spark, Flink etc.

Even if we have a big data solution to process the high quantity of raw data, it is not the efficiency and performance of the solution that determines the quality of the knowledge extracted but it depends on the quality of the data as well. The raw data likely to be imperfect and may contain noise, irrelevant information, missing values etc. It is well known that low quality data will lead to low quality knowledge [9]. Hence data cleaning is the major step to be performed before we continue with data mining algorithms to make sure that we are using a suitable and relevant data set.

Data cleaning has 2 parts. First part is data preparation and second part is data reduction techniques.

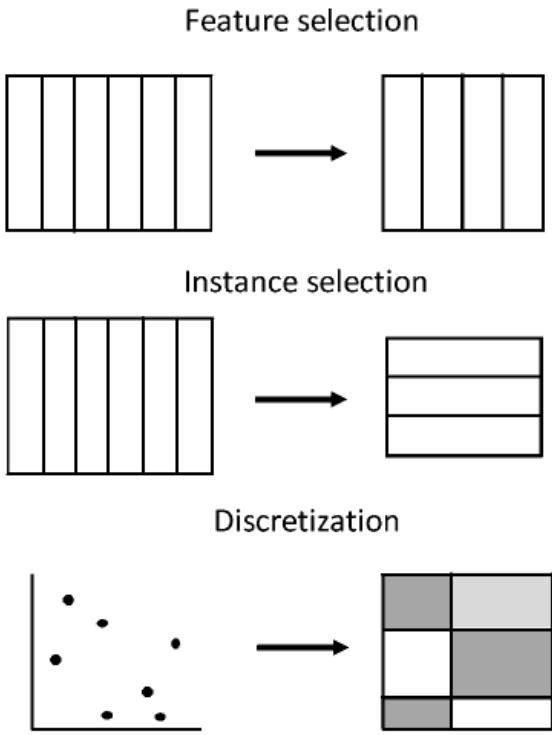
**3.0.1 Data Preparation.** The data going to the analytics model should be clean and noise free. Hence data preparation part includes tasks like data cleaning, data normalization, data transformation, missing value imputation, data integration and Noise identification. Figure 5 shows the data preparations tasks [9].



**Figure 5: Data Preprocessing and preparation tasks [9]**

**3.0.2 Data Reduction Techniques.** To reduce the dimensionality problem and the computational cost, because of a large number of variables and instances in the data set, we try to gather only the required set of quality data. Data reduction techniques include feature selection, instance selection and discretization. Figure 6 shows data reduction techniques [9].

With respect to our chosen data set, data reduction techniques were already applied to the raw data and 1000 customer records and 21 variables were shortlisted. All these variables are either categorical (like Account-Balance, Previous-credit, purpose etc) or continuous (Duration-of-credit, Installment-percent, dependents). As part of data preparation for our analysis, we transform the values of categorical variable's from string to scores (numerical values). For example, the variable "creditability" got 2 values - good and bad. After transformation process, "good" get replaced by "1" and "bad" get replaced by "0". Likewise, we gave scores for the values of the variables, Foreign-Worker, Telephone, Previous-Credit, Purpose, Sex-MaritalStatus, Guarantors and Type-apartment.



**Figure 6: Data Reduction Approaches [9]**

## 4 DATA ANALYSIS

Big data analysis is the process of obtaining knowledge by analyzing and understanding hidden patterns, market trends, unknown correlations, customer preferences and other relevant information from large and varied datasets [21]. Big data analytics methods include exploratory analysis, data mining, predictive analytics, machine learning, deep learning etc. The results of the analysis can be visualized using tools like Tableau, Infogram, Plotly etc or by using Python scripts. This project utilizes methods like exploratory analysis, predictive analysis, machine learning algorithms and visualizations of results using Python scripts. Python codes for all these analysis methods are given in appendix A.3.

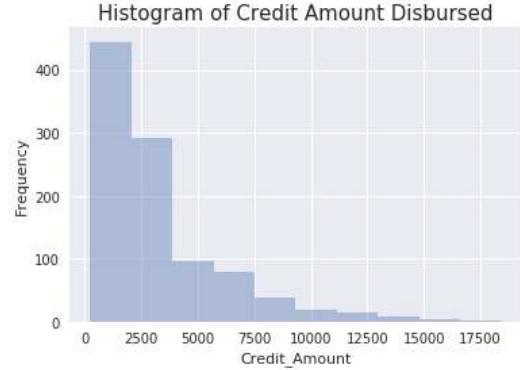
### 4.1 Exploratory Analysis

Exploratory analysis is basically to explore the data and understand what it actually contains. It is an approach to summarize the general characteristics of the data set before we attempt to model it. Statistical methods or direct visualizations can help in data exploration [32].

**4.1.1 Direct Visualization.** After data preprocessing, our dataset includes 1000 customer records with 20 feature variable and 1 class variable. Feature variable values can be visualized to understand the characteristics and how they are related to each other - proportionally or inversely.

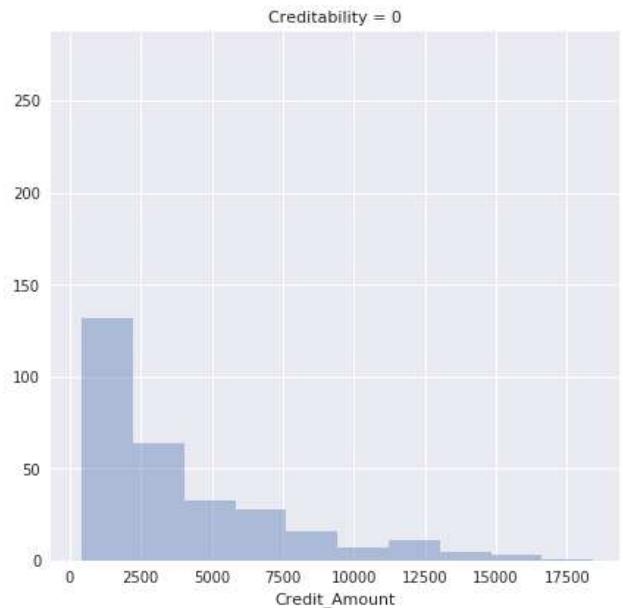
Figure 7 shows the histogram of credit amount disbursed with respect to frequency. From this diagram, we understand that most

of the customers are requested for loans for up to 2500 German Marks. The number of customers decreases as the loan amount increases. And very few customers fall under the loan amount category over 10000 German Marks.



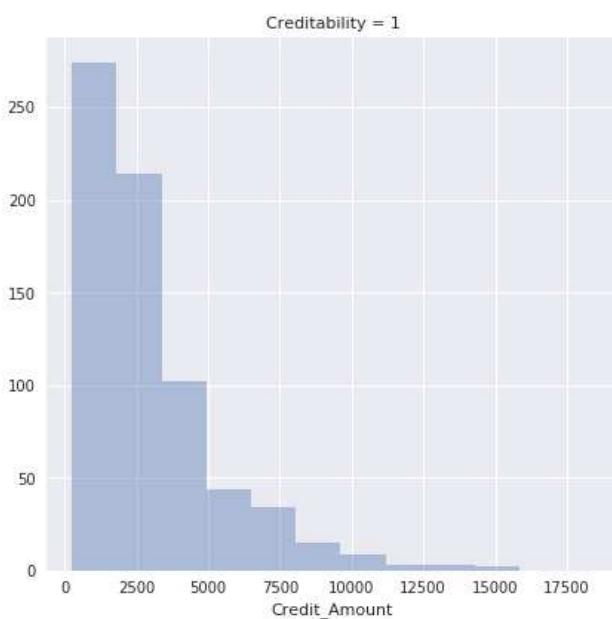
**Figure 7: Credit amount vs. Frequency**

Figure 8 and Figure 9 shows the credit amount availed by bad customers and good customers respectively. The trend is almost the same. Maximum customers from both the classes fall under the category of up to 2500 German Marks. But there is a noticeable difference in the number of customers under 12500 range. Bad rated customers are more in this category.

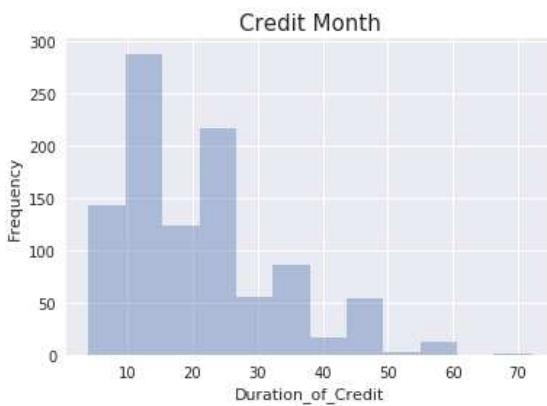


**Figure 8: Credit amount vs. bad customers**

Figure 10 shows the duration of credit in months vs. number of customers. From this graph, we can understand that maximum number of customers opted for 10 to 15 months duration.



**Figure 9: Credit amount vs. good customers**



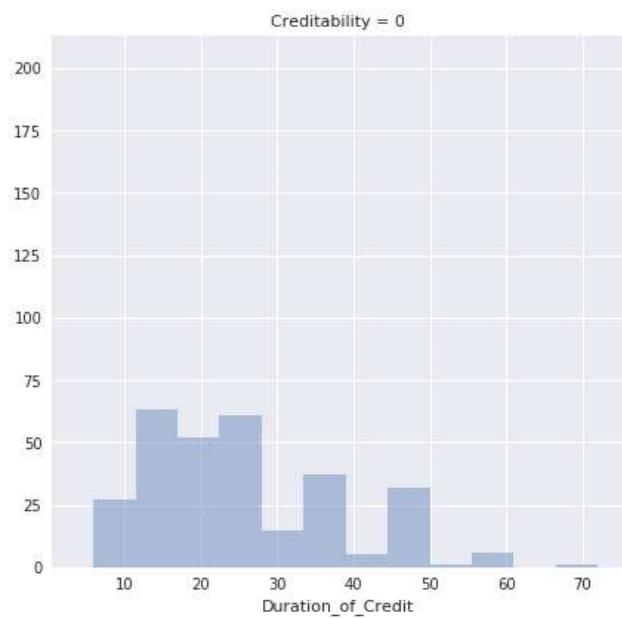
**Figure 10: Duration of credit in months vs. frequency**

Figure 11 and Figure 12 shows the duration of credit in months vs number of bad customers and good customers respectively. It shows that there is not much difference in the trend.

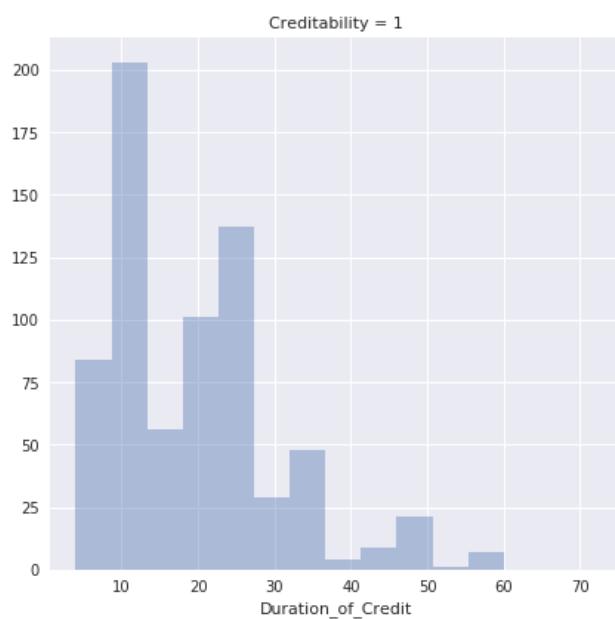
Figure 13 shows how customers are scattered with respect to age. Most of the borrowers fall under the age group of 23 to 28.

**4.1.2 Data Classification.** We have one class variable "Creditability" to classify the customers based on the bank's opinion on the actual applicants. We could extract this class information from dataset using PySpark Python script "GroupBy". Figure 14 shows the output of the script.

Customers in our dataset are classified into 2 classes - Good (1 = Creditworthy) and Bad (0 = Uncreditworthy). We have 700 customers in the Good class and 300 customers in the Bad class.



**Figure 11: Duration of credit in months vs. bad customers**



**Figure 12: Duration of credit in months vs. good customers**

We divide our dataset of 1000 customer records randomly into 2 parts. First part is the training dataset with 700 customer records and second part is the test dataset of 300 customer records.

**4.1.3 Interquartile Range.** Interquartile Range is a statistical method to measure the variability of the data. This will be applicable only for the continuous variables (Credit-amount, Duration of credit and Age). The rank-ordered data will be divided into 4 equal parts

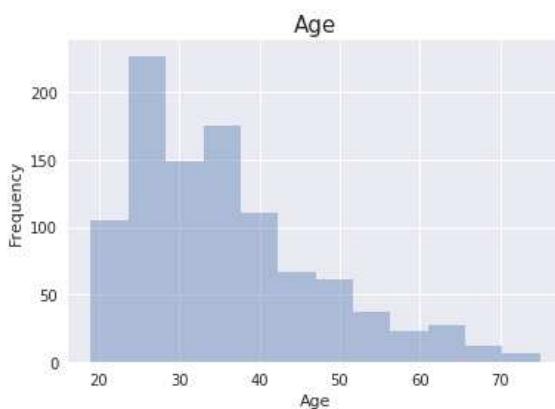


Figure 13: Age vs. frequency

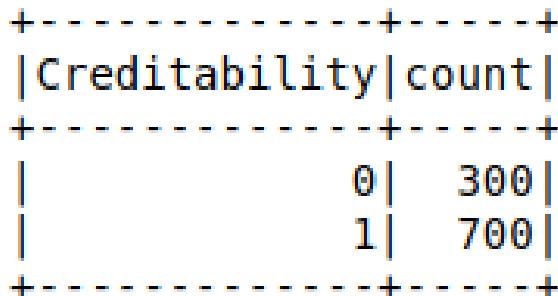


Figure 14: Data classification

called quartiles. Values are called the First (Q1) Second (Q2) and Third (Q3) quartiles. Q2 is the Median value of the dataset [29].

We used pandas quantile function to extract this information for all the continuous variables.

Figure 15 shows the variability of Credit-Amount.

	Min	1st Qu	Median	Mean	3rd Qu	Max
0	250	1365.5	2319.5	3271.248	3972.25	18424

Figure 15: Variability in Credit-Amount

Figure 16 shows the variability of Duration of credit.

	Min	1st Qu	Median	Mean	3rd Qu	Max
0	4	12.0	18.0	20.903	24.0	72

Figure 16: Variability of Duration of credit

Figure 17 shows the variability of Age.

	Min	1st Qu	Median	Mean	3rd Qu	Max
0	19	27.0	33.0	35.542	42.0	75

Figure 17: Variability of Age

**4.1.4 Cross-Tabulation.** Cross-Tabulation is a statistical method used to compare the relationship between categorical variables. In our scenario, we examine the relationship of the categorical variables with the class variable "creditability". We create a *contingency table* which displays the frequency of categorical variables with respect to the class [35].

	Sex_MaritalStatus				
Creditability	1	2	3	4	Row Total
good	30 0.6	201 0.6	402 0.7	67 0.7	700
bad	20 0.4	109 0.4	146 0.3	25 0.3	300
Column Total	50 0.0	310 0.3	548 0.5	92 0.1	1000

Figure 18: Contingency table of sex-marital status [15]

Figure 18 shows the contingency table created for the variable sex-marital status against class. It shows the number of good and bad customers distributed among the 4 categories of the variable sex-marital status. Category "male: married/widowed" has the maximum number of Good customers. Contingency tables are used to create the Chi-square values.

**4.1.5 Test of Independence.** We need to identify the features that are closely related to the class/credit rating to build a predictive model. We do a test of independence on all our feature variables to identify the ones to be selected for data modeling. The method we use to do the Test of Independence is the Chi-squared test. The output of the Chi-squared test is the input to the Logistical Regression Algorithm. Variables which are not related to the class variable will be discarded from further analysis of Logical Regression.

*Pearson's Chi-squared test:* Chi-squared test is used to determine the significant difference between expected values and observed values in one or more categories. There are 2 types of Chi-squared test - Goodness of Fit and Test for Independence.

We use the second method - *Test of Independence*. It compares 2 variables in a contingency table to check if they are related. In other words, it examines if the distributions of categorical variables are different from one another.

If the calculated value is small that means, the variables are related. If the value is large that means, the data is not related and not fit for analysis [28].

*p-value:* p-value is the probability value that, when the null hypothesis is true, the chi-square value will be greater than the empirical value of the data. There is a p-value distribution chart available where it is calculated against the significance value, degrees of freedom and chi-square test value [13].

*Degrees of freedom:* Degrees of freedom is the number of scores that can be varied. It is calculated using the formula,

$$\text{Degrees of freedom} = (r - 1) * (c - 1) \quad (1)$$

The calculated values are shown in figure 19.

	All	Chi^2	D.F	PValues
0	Account_Balance	123.720944	3	0.000000e+00
1	Duration_of_Credit	78.886937	32	7.784572e-06
2	Previous_Credit	61.691397	4	1.279199e-12
3	Purpose	33.356447	9	1.157491e-04
4	Credit_Amount	931.746032	922	4.045155e-01
5	Value_Savings_Stocks	36.098928	4	2.761214e-07
6	employment	18.368274	4	1.045452e-03
7	Instalment_percent	5.476792	3	1.400333e-01
8	Sex_MaritalStatus	9.605214	3	2.223801e-02
9	Guarantors	6.645367	2	3.605595e-02
10	Duration_address	0.749296	3	8.615521e-01
11	asset	23.719551	3	2.858442e-05
12	Age	57.626982	52	2.749531e-01
13	Concurrent_Credits	12.839188	2	1.629318e-03
14	Type_apartment	18.674005	2	8.810311e-05
15	No_of_Credits	2.671198	3	4.451441e-01
16	Occupation	1.885156	3	5.965816e-01
17	dependents	0.009089	1	9.240463e-01
18	Telephone	1.329783	1	2.488438e-01
19	Foreign_Worker	6.737044	1	9.443096e-03

Figure 19: Chi-square, df and p values

## 5 MODELS

Predictive models can be created using different Machine Learning algorithms such as Logistical Regression, Decision Trees, Random Forest etc. Machine learning algorithms generate models from the training data and tested against the test data to estimate the accuracy level. Before building predictive models, there are few baseline models can be created to compare and see what improvements we are actually trying to achieve. By comparing the accuracies of different predictive models against the base models, we can come up with the best model for that particular problem. The best model is saved for the future predictions on new datasets.

### 5.1 Baseline Models

Baseline models use simple summary statistics. In classification problems like our scenario, baseline models are created based on the class values. As mentioned in the data classification section 4.1.2, our total list of 1000 customer records are divided into training

Table 2: Baseline Model 1

		Good
Good	Bad	
210	70	

Table 3: Baseline Model 2

		Bad
Good	Bad	
210	70	

dataset and test dataset. Training dataset has 700 customer records and test dataset has 300 customer records. For the baseline models, we evaluate the test data of 300 customer records.

In this project we create 2 baseline models.

*Baseline Model 1:* In this model, we assume all the input test customer records (300 customer records) belongs to the "Good" class. Since out of 1000 customers, 700 falls under "Good" class, we assume among the 300 customers in test dataset 70% will fall under "Good" class and rest in "Bad" class, which means this baseline model holds 70% accuracy.

Table 2 shows the assumption in baseline model 1.

*Baseline Model 2:* In this model, we assume all the input test customer records (300 customer records) belongs to the "Bad" class. Since out of 1000 customers, 300 falls under "Bad" class, we assume among the 300 customers in test dataset 30% will fall under "Bad" class and rest in "Good" class, which means this baseline model holds 30% accuracy.

Table 3 shows the assumption in baseline model 2.

### 5.2 Decision Tree Model

To build this model, we use the machine learning algorithm - Decision Tree which is explained in section 1.2.6. PySpark's class "DecisionTreeClassifier" is used to build different Decision Tree models from training data based on different tree attributes like MaxBins, Maxdepth, Impurity etc. Impurity measures are calculated internally by this classifier to identify the root node and other internal nodes. Gini Index is the method opted in our project.

Formula to calculate Gini Index is,

$$GiniIndex = \sum_{i=1}^C f_i(1 - f_i) \quad (2)$$

We created 2 Decision Tree models to compare the accuracy.

*Decision Tree with maxDepth None:* In this model, we set the maxDepth value of the Tree to None and we calculated the accuracy using PySpark's "MulticlassClassificationEvaluator". In this case, the tree can become arbitrarily deep and complex and more chances of overfitting issues.

The accuracy of the output of this model is 0.679. Maximum number of Bins are 32. Depth is None.

*Decision Tree after adjusting the attribute values:* In this model, we set the maxDepth value to 6 and maxBins value to 20. We used

the same PySpark's "MulticlassClassificationEvaluator" to calculate the accuracy. Since we have limited the maxDepth and maxBin values, the overfitting issues decreases.

- The accuracy of the output of this model is 0.716
- Number of Bins are 20
- Depth is 6

### 5.3 Random Forest

Random Forest Machine Learning algorithm which is explained in the section 1.2.7 is used to build Random Forest model. We use PySpark class "RandomForestClassifier" to generate the model from training data. We build 2 Random Forest models one with default attribute and another one with chosen attribute values.

*Random Forest with Default Settings:* In this case, the attributes of the Tree are selected by the "RandomForestClassifier" itself internally and accuracy of the model is calculated based on that.

The accuracy of the output of this model is 0.756. Maximum number of Bins are 32. Maximum Depth is 5. Maximum number of Trees are 20.

*Tuning Random Forest with cross-validator:* In this case, we tune the Random Forest model by trying different attribute values for tree attributes - maxDepth, maxBin and numTrees. We can provide multiple values for each attribute. We provided 3 values for maxDepth, 2 values for maxBins and 3 values for numTrees. We will start with some random values for these attributes.

We use *cross-validation* techniques in this type of Random Forest model to get the best model. PySpark "CrossValidator" will analyze the values of the attributes. In this scenario, the "CrossValidator" will choose 3 values of attributes from  $3 * 2 * 3$  values. It will then try different combinations of the attribute values internally and finally, the model will get tuned to a final set of attributes which derive the best model with maximum accuracy.

- The accuracy of the output of this model is 0.779
- Number of Trees are 100

As we identified the best model with maximum accuracy is the Random Forest model, we passed the actual dataset to this model and received an accuracy of 0.845

## 6 RESULTS

Now we have all the desired models created which can predict the class of a new customer. We can compare and analyze the outputs of each of these models and conclude with the best model. We can analyze the results based on accuracy and mean penalty matrix.

### 6.1 Prediction matrix

Prediction matrix can be extracted using the "groupby" option in PySpark. Figure 20 and Figure 21 shows the prediction matrix of Decision Tree and Random Forest respectively. Decision Tree has got 219 right predictions and Random Forest has got 240 right predictions out of 300 customer records.

### 6.2 Feature Importance

Feature Importance is the list of important predictors that are the top contributed variables towards building the predictive model. Normally the variable with maximum dependency would be treated

**prediction 0.0 1.0**

**label**

0.0	35	55
1.0	34	184

Figure 20: Prediction matrix - Decision Tree

**prediction 0.0 1.0**

**label**

0.0	40	50
1.0	18	200

Figure 21: Prediction matrix - Random Forest

as the root node by the algorithm. We could calculate the feature importance only for the Random forest algorithm by using the class "bestModel.featureImportances". Figure 22 shows the list of predictors. We could see that the variable "Account Balance" contributes maximum to the predictions.

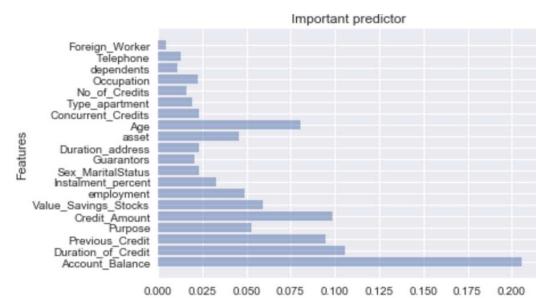


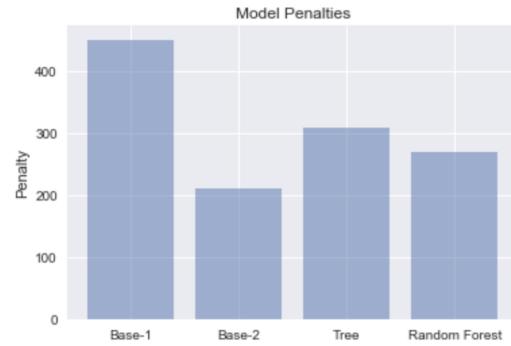
Figure 22: Random Forest Important Predictors

### 6.3 Model Accuracy Comparison

Figure 23 shows the accuracy of different predictive models that we created. We plotted the output of "MulticlassClassificationEvaluator" for Decision Tree and Random Forest. We can understand from the graph that baseline model 2 has got the least accuracy and Random Forest has got the most.



**Figure 23: Model accuracy comparison**



**Figure 24: Model penalty comparison**

**Table 4: Penalty Matrix [15]**

Actual	Predicted 'Good'	Predicted 'Bad'
Good	0	1
Bad	5	0

#### 6.4 Penalty Matrix

One important aspect to consider while choosing a predictive model is the accuracy. When considering the actual goal of this project, the model should be apt to minimize the risks and to maximize the profit. The model should ensure good prediction accuracy to achieve the goal.

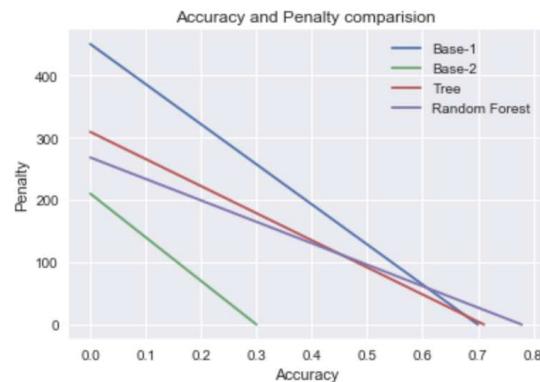
A penalty matrix is defined to calculate the loss to the bank. Penalty will be applied to each misclassifications and penalty value differs for wrong classifications - 'good as bad' and 'bad as good'. As discussed in the project goals section 1.1, approving loan for an uncreditworthy customer will end up in more financial loss for the bank and accordingly is a greater risk. Hence classifying a bad customer wrongly as good customer will have more penalty.

Table 4 shows the penalty matrix. For right predictions penalty is 0. If a good customer predicted as bad, the penalty is 1 and if a bad customer predicted as good, the penalty is 5. The sum of the penalty values multiplied with the respective number of misclassified customers will provide the total amount of loss / penalty.

Figure 24 shows the penalty comparison of different predictive models. Baseline model 1 has more chances of predicting bad customers as good because it blindly assumes that, all the incoming customers are good. Hence it has got more penalty value. Baseline model 2 has got least chances of classifying bad customers as good because it assumes all incoming customers are bad. Hence baseline model 2 has minimum penalty.

#### 6.5 Accuracy and Penalty Comparison

Figure 25 shows the accuracy and penalty comparison for all the 4 models. Random Forest is the most accurate model and with minimal penalty. Hence Random Forest is the best model out of all.



**Figure 25: Model accuracy and penalty comparison**

## 7 DISCUSSION

We have built 4 predictive models. Baseline model 1 and 2, Decision Tree and Random Forest. We did a small study on Logistical Regression model as well. There are many other machine learning algorithms available which are suitable for classification analysis. Current analysis uses only 20 feature variables and 1000 customer records to populate the predictive models. In predictive analysis, the bigger the training dataset, the better the outcome. Current analysis can be extended to really big data with more feature variables customer records and also data from multiple years. Data processing can be done using distributed big data processing systems available today for better accuracy. Unfortunately, such a large data is not publicly available for studies in finance area right now. Hence we tried big data technologies in a comparatively smaller dataset.

## 8 CONCLUSION

Out of 4 predictive models created, Random Forest has the maximum accuracy in classifying the customers in the right class. Even if it gives an accuracy of around 85% it is not an error free model. There are 15% chances for misclassification. The size of the dataset that we considered to develop this model may have a direct impact. If we can train the model with a larger data set with tens of thousands of customer records and feature variables, the accuracy

may increase close to 100%. There might be other more advanced machine learning algorithms and tools coming up to explore the chances of increasing the overall accuracy of the predictive models in common.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Juliette Zurick, Miao Jiang and Saber Sheybani Moghadam for their suggestions and support to complete this project and report.

## REFERENCES

- [1] askubuntu. 2015. How do I get pyspark on Ubuntu? Web page. (June 2015). <https://askubuntu.com/questions/635265/how-do-i-get-pyspark-on-ubuntu>
- [2] askubuntu. 2016. how to install numpy for python3. Web page. (April 2016). <https://askubuntu.com/questions/765494/how-to-install-numpy-for-python3/765510>
- [3] askubuntu. 2016. Unable to install matplotlib using pip in Ubuntu 16.04. Web page. (June 2016). <https://askubuntu.com/questions/791673/unable-to-install-matplotlib-using-pip-in-ubuntu-16-04>
- [4] askubuntu. 2017. How do I install Python 3.6 using apt-get? Web page. (November 2017). <https://askubuntu.com/questions/865554/how-do-i-install-python-3-6-using-apt-get>
- [5] Charles Bochet. 2017. Get Started with PySpark and Jupyter Notebook in 3 Minutes. Web page. (May 2017). <https://blog.sicara.com/get-started-pyspark-jupyter-guide-tutorial-ac2fe84f594f>
- [6] Matplotlib development team. 2017. Matplotlib Introduction. Web page. (October 2017). <https://matplotlib.org/users/intro.html>
- [7] dezyre.com. 2017. PySpark Tutorial-Learn to use Apache Spark with Python. Web page. (September 2017). <https://www.dezyre.com/apache-spark-tutorial/pyspark-tutorial>
- [8] Dhanya. 2017. code. Web page. (November 2017). <https://github.com/bigdata-i523/hid328/tree/master/project/code>
- [9] Salvador Garcia, Sergio Ramirez-Gallego, Julian Luengo, Jose Manuel Benitez, and Francisco Herrera. 2016. Big data preprocessing: methods and prospects. Web page. (September 2016). [https://bdataalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0](https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0)
- [10] Dr. Hans Hofmann. 1994. Statlog (German Credit Data) Data Set. Web page. (November 1994). <https://archive.ics.uci.edu/ml/datasets/Statlog+German+Credit+Data%29>
- [11] Katharine Jarmul. 2016. INTRODUCTION TO DATA SCIENCE: HOW TO fIBIG DATAfI WITH PYTHON. Web page. (October 2016). <http://dataconomy.com/2016/10/big-data-python/>
- [12] Dhanya Mathew. 2017. dataset in excel format. Web page. (November 2017). <https://github.com/bigdata-i523/hid328/tree/master/project>
- [13] medcalc. 2015. Values of the Chi-squared distribution. Web page. (April 2015). <https://www.medcalc.org/manual/chi-square-table.php>
- [14] Trevor Nath. 2015. How Big Data Has Changed Finance. (April 2015). <http://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp>
- [15] PennState Eberly College of Science. 2016. Analysis of German Credit Data. Web page. (September 2016). <https://onlinecourses.science.psu.edu/stat857/node/215>
- [16] pandas. 2017. Installation. Web page. (June 2017). <https://pandas.pydata.org/pandas-docs/stable/install.html>
- [17] pandas.pydata.org. 2017. pandas: powerful Python data analysis toolkit. Web page. (October 2017). <https://pandas.pydata.org/pandas-docs/stable/>
- [18] pip.pypa.io. 2016. Installation. Web page. (July 2016). <https://pip.pypa.io/en/stable/installing/>
- [19] Saimadhu Polamuri. 2017. HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING. Web page. (May 2017). <https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learning/>
- [20] rosehosting.com. 2017. How to Install Jupyter on an Ubuntu 16.04. Web page. (February 2017). <https://www.rosehosting.com/blog/how-to-install-jupyter-on-an-ubuntu-16-04-vps/>
- [21] Margaret Rouse. 2017. big data analytics. Webpage. (July 2017). <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>
- [22] Fabrizio Sarrocco, Vincenzo Morabito, and Gregor Meyer. 2016. Exploring Next Generation Financial Services: The Big Data Revolution. (2016). [https://www.accenture.com/t20170314T051509\\_w\\_/\\_nl-en/\\_acnmedia/PDF-20/Accenture-Next-Generation-Financial.pdf](https://www.accenture.com/t20170314T051509_w_/_nl-en/_acnmedia/PDF-20/Accenture-Next-Generation-Financial.pdf)
- [23] sas. 2017. Machine Learning What it is and why it matters. Web page. (June 2017). [https://www.sas.com/en\\_us/insights/analytics/machine-learning.html](https://www.sas.com/en_us/insights/analytics/machine-learning.html)
- [24] Rahul Saxena. 2017. Introduction to Decision Tree Algorithm. Web page. (January 2017). <https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>
- [25] Luca Scagliarini. 2017. What is Machine Learning? A definition. Web page. (July 2017). <http://www.expertsystem.com/machine-learning-definition/>
- [26] seaborn. 2017. Seaborn Installing and getting started. Web page. (June 2017). <https://seaborn.pydata.org/installing.html>
- [27] Sabeer Shaikh. 2016. Why Python is important for big data and analytics applications? Web page. (April 2016). <https://www.eduonix.com/blog/bigdata-and-hadoop/python-important-big-data-analytics-applications/>
- [28] statisticshowto.com. 2016. Chi-Square Statistic: How to Calculate It - Distribution. Web page. (June 2016). <http://www.statisticshowto.com/probability-and-statistics/chi-square/>
- [29] Stat Trek. 2017. Statistics and Probability Dictionary. Web page. (November 2017). <http://stattrek.com/statistics/dictionary.aspx?definition=Interquartile%20range>
- [30] Michael Waskom. 2017. seaborn: statistical data visualization. Web page. (October 2017). <https://seaborn.pydata.org/>
- [31] Wiki. 2017. Big data. Web page. (Oct 2017). [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data)
- [32] wiki. 2017. Exploratory data analysis. Web page. (October 2017). [https://en.wikipedia.org/wiki/Exploratory\\_data\\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)
- [33] wiki. 2017. Machine learning. Web page. (October 2017). [https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)
- [34] wiki. 2017. NumPy. Web page. (October 2017). <https://en.wikipedia.org/wiki/NumPy>
- [35] Yolanda Williams. 2015. Cross Tabulation: Definition & Examples. Web page. (June 2015). <http://study.com/academy/lesson/cross-tabulation-definition-examples-quiz.html>

## A PROJECT REFERENCES

All project related documents are available in the github repository i523/hid328/project: <https://github.com/bigdata-i523/hid328/tree/master/project> [12].

### A.1 Makefile

Make file is created assuming that the target system has Ubuntu OS and Python3 installed already. This can be executed from terminal window from folder i523/hid328/project/code using command "make run". Makefile is available in the github repository i523/hid328/project/code [8].

### A.2 Data Set

"credit-data.csv" is available in Google Drive /project-data/hid328/.

### A.3 Project Code

Project code is available in the Jupyter notebook "project.ipynb" in the github repository i523/hid328/project/code [8].

```
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
```

```
=====
[2017-12-10 13.51.57] 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 1.8s.
```

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

```
=====
name: Dhanya Mathew
hid: 328
paper1: Nov 2 17 100%
paper2: Nov 6 17 100%
project: Dec 4 17 100%
```

```
=====
yamlcheck
```

```
wordcount
```

---

```
11  
wc 328 project 11 5983 report.tex  
wc 328 project 11 6151 report.pdf  
wc 328 project 11 1015 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
```

---

```
44: \subsection{Project Goals}\label{Project goals}  
50: \subsection{Methods and Technologies Involved}\label{Methods and  
Technologies Involved}  
56: Overall workflow of the project is shown in Figure  
    \ref{fig:Figure1}.  
57: We have taken a sample data set of loan applications received by a  
    bank. We explored the data and the requirements of the bank and  
    based on that set the project goals as discussed in the section  
    \ref{Project goals}, before starting the project. In the real  
    scenarios, we will not be able to apply analytical methods  
    directly on the raw data as it likely be imperfect and containing  
    irrelevant information. Hence we do data cleaning (data
```

preprocessing) as the first step. Data cleaning is done using PySpark. The cleaned data has 1000 customer records with 1 classifier and 20 feature variables.

```

59: \begin{figure}[htb]
61: \includegraphics[width=1.0\columnwidth]{images/Figure1.png}
64: \label{fig:Figure1}
106: \subsubsection{Decision Tree}\label{Decision Tree}
108: Decision Tree is a supervised machine learning algorithm used to solve both classification and regression problems. In Decision Tree, a trained model with a set of rules will be created based on the training data. The target class or value of a test/new data set will be predicted based on this training rules. Decision Tree algorithm is simple to understand as it uses a tree model representation to solve the problem. It starts from a root node and continues with other decision nodes. Each internal decision nodes corresponds to the feature variables and each leaf nodes corresponds to the class label \cite{decision-tree}. Figure \ref{fig:Figure2} shows the decision tree classifier.
  
```

```

111: \begin{figure}[htb]
113: \includegraphics[width=1.0\columnwidth]{images/Figure2.png}
116: \label{fig:Figure2}
123: \subsubsection{Random Forest}\label{Random Forest}
125: Like Decision Tree, Random Forest algorithm also can be used for classification as well as regression problems. It is a supervised machine learning algorithm. It uses decision tree concept as well but there will be more than one trees in a Random Forest. As the number of trees increases in the Random forest, the accuracy of the prediction also will increase accordingly. Random forest algorithm can handle missing values in the data. Also with more trees in the forest, overfitting issues will not occur in Random Forest algorithm \cite{random-forest}. Figure \ref{fig:Figure3} shows the Random Forest model.
  
```

```

127: \begin{figure}[htb]
129: \includegraphics[width=1.0\columnwidth]{images/Figure3.png}
132: \label{fig:Figure3}
135: Random Forest algorithm progresses via 2 stages - Random Forest creation and Perform prediction. To create the Random Forest, we select a random number of feature variables from the total list of feature variables in the training data and create a Decision Tree out of it. We repeat this process to create desired number of trees. These randomly created trees will form a Random Forest. Figure \ref{fig:Figure4} shows how random forest algorithm works.
  
```

```

137: \begin{figure}[htb]
139: \includegraphics[width=1.0\columnwidth]{images/Figure4.png}
142: \label{fig:Figure4}
150: Technologies used in this project are discussed in detail in
  
```

- section \ref{Methods and Technologies Involved}. The installation commands on Ubuntu 16.04 OS for each of these technologies are given in this section. Installations can be done from the terminal window.
- 163: All these installation steps are included in a make file referred in appendix \ref{Makefile}.
- 169: Dataset includes 1000 customer records with 20 feature variables and a class variable. In the class variable, the actual class of the customer is specified - good or bad. The complete list of data set variables and their description is given in Table \ref{tab:table1}.
- 171: \begin{table}
- 174: \label{tab:table1}
- 213: The data going to the analytics model should be clean and noise free. Hence data preparation part includes tasks like data cleaning, data normalization, data transformation, missing value imputation, data integration and Noise identification. Figure \ref{fig:Figure5} shows the data preparations tasks \cite{preprocessing}.
- 215: \begin{figure}[htb]
- 217: \includegraphics[width=1.0\columnwidth]{images/Figure5.png}
- 220: \label{fig:Figure5}
- 225: To reduce the dimensionality problem and the computational cost, because of a large number of variables and instances in the data set, we try to gather only the required set of quality data. Data reduction techniques include feature selection, instance selection and discretization. Figure \ref{fig:Figure6} shows data reduction techniques \cite{preprocessing}.
- 227: \begin{figure}[htb]
- 229: \includegraphics[width=1.0\columnwidth]{images/Figure6.png}
- 232: \label{fig:Figure6}
- 240: Big data analysis is the process of obtaining knowledge by analyzing and understanding hidden patterns, market trends, unknown correlations, customer preferences and other relevant information from large and varied datasets \cite{bigdata-analytics}. Big data analytics methods include exploratory analysis, data mining, predictive analytics, machine learning, deep learning etc. The results of the analysis can be visualized using tools like Tableau, Infogram, Plotly etc or by using Python scripts. This project utilizes methods like exploratory analysis, predictive analysis, machine learning algorithms and visualizations of results using Python scripts. Python codes for all these analysis methods are given in appendix \ref{Project Code}.
- 250: Figure \ref{fig:Figure7} shows the histogram of credit amount disbursed with respect to frequency. From this diagram, we

understand that most of the customers are requested for loans for up to 2500 German Marks. The number of customers decreases as the loan amount increases. And very few customers fall under the loan amount category over 10000 German Marks.

```
252: \begin{figure}[htb]
254: \includegraphics[width=1.0\columnwidth]{images/Figure7.png}
256: \label{fig:Figure7}
259: Figure \ref{fig:Figure8} and Figure \ref{fig:Figure9} shows the credit amount availed by bad customers and good customers respectively. The trend is almost the same. Maximum customers from both the classes fall under the category of up to 2500 German Marks. But there is a noticeable difference in the number of customers under 12500 range. Bad rated customers are more in this category.
261: \begin{figure}[htb]
263: \includegraphics[width=1.0\columnwidth]{images/Figure8.png}
265: \label{fig:Figure8}
268: \begin{figure}[htb]
270: \includegraphics[width=1.0\columnwidth]{images/Figure9.png}
272: \label{fig:Figure9}
275: Figure \ref{fig:Figure10} shows the duration of credit in months vs. number of customers. From this graph, we can understand that maximum number of customers opted for 10 to 15 months duration.
277: \begin{figure}[htb]
279: \includegraphics[width=1.0\columnwidth]{images/Figure10.png}
281: \label{fig:Figure10}
284: Figure \ref{fig:Figure11} and Figure \ref{fig:Figure12} shows the duration of credit in months vs number of bad customers and good customers respectively. It shows that there is not much difference in the trend.
286: \begin{figure}[htb]
288: \includegraphics[width=1.0\columnwidth]{images/Figure11.png}
290: \label{fig:Figure11}
293: \begin{figure}[htb]
295: \includegraphics[width=1.0\columnwidth]{images/Figure12.png}
297: \label{fig:Figure12}
300: Figure \ref{fig:Figure13} shows how customers are scattered with respect to age. Most of the borrowers fall under the age group of 23 to 28.
302: \begin{figure}[htb]
304: \includegraphics[width=1.0\columnwidth]{images/Figure13.png}
306: \label{fig:Figure13}
309: \subsubsection{Data Classification}\label{Data Classification}
311: We have one class variable ''Creditability'' to classify the customers based on the bank's opinion on the actual applicants. We could extract this class information from dataset using
```

PySpark Python script ''GroupBy''. Figure \ref{fig:Figure14} shows the output of the script.

```

313: \begin{figure}[htb]
315: \includegraphics[width=1.0\columnwidth]{images/Figure14.png}
317: \label{fig:Figure14}
329: Figure \ref{fig:Figure15} shows the variability of Credit-Amount.
331: \begin{figure}[htb]
333: \includegraphics[width=1.0\columnwidth]{images/Figure15.png}
335: \label{fig:Figure15}
338: Figure \ref{fig:Figure16} shows the variability of Duration of credit.
340: \begin{figure}[htb]
342: \includegraphics[width=1.0\columnwidth]{images/Figure16.png}
344: \label{fig:Figure16}
347: Figure \ref{fig:Figure17} shows the variability of Age.
349: \begin{figure}[htb]
351: \includegraphics[width=1.0\columnwidth]{images/Figure17.png}
353: \label{fig:Figure17}
360: \begin{figure}[htb]
362: \includegraphics[width=1.0\columnwidth]{images/Figure18.png}
365: \label{fig:Figure18}
368: Figure \ref{fig:Figure18} shows the contingency table created for the variable sex-marital status against class. It shows the number of good and bad customers distributed among the 4 categories of the variable sex-marital status. Category ''male: married/widowed'' has the maximum number of Good customers. Contingency tables are used to create the Chi-square values.
389: The calculated values are shown in figure \ref{fig:Figure19}.
391: \begin{figure}[htb]
393: \includegraphics[width=1.0\columnwidth]{images/Figure19.png}
395: \label{fig:Figure19}
404: Baseline models use simple summary statistics. In classification problems like our scenario, baseline models are created based on the class values. As mentioned in the data classification section \ref{Data Classification}, our total list of 1000 customer records are divided into training dataset and test dataset. Training dataset has 700 customer records and test dataset has 300 customer records. For the baseline models, we evaluate the test data of 300 customer records.
410: \begin{table}
412: \label{tab:table2}
423: Table \ref{tab:table2} shows the assumption in baseline model 1.
427: \begin{table}
429: \label{tab:table3}
440: Table \ref{tab:table3} shows the assumption in baseline model 2.
444: To build this model, we use the machine learning algorithm -

```

Decision Tree which is explained in section \ref{Decision Tree}. PySpark's class ''DecisionTreeClassifier'' is used to build different Decision Tree models from training data based on different tree attributes like MaxBins, Maxdepth, Impurity etc. Impurity measures are calculated internally by this classifier to identify the root node and other internal nodes. Gini Index is the method opted in our project.

- 468: Random Forest Machine Learning algorithm which is explained in the section \ref{Random Forest} is used to build Random Forest model. We use PySpark class ''RandomForestClassifier'' to generate the model from training data. We build 2 Random Forest models one with default attribute and another one with chosen attribute values.
- 492: Prediction matrix can be extracted using the ''groupby'' option in PySpark. Figure \ref{fig:Figure20} and Figure \ref{fig:Figure21} shows the prediction matrix of Decision Tree and Random Forest respectively. Decision Tree has got 219 right predictions and Random Forest has got 240 right predictions out of 300 customer records.
- 494: \begin{figure}[htb]  
496: \includegraphics[width=1.0\columnwidth]{images/Figure20.png}  
498: \label{fig:Figure20}  
501: \begin{figure}[htb]  
503: \includegraphics[width=1.0\columnwidth]{images/Figure21.png}  
505: \label{fig:Figure21}  
510: Feature Importance is the list of important predictors that are the top contributed variables towards building the predictive model. Normally the variable with maximum dependency would be treated as the root node by the algorithm. We could calculate the feature importance only for the Random forest algorithm by using the class ''bestModel.featureImportances''. Figure \ref{fig:Figure22} shows the list of predictors. We could see that the variable ''Account Balance'' contributes maximum to the predictions.  
512: \begin{figure}[htb]  
514: \includegraphics[width=1.0\columnwidth]{images/Figure22.png}  
516: \label{fig:Figure22}  
521: Figure \ref{fig:Figure23} shows the accuracy of different predictive models that we created. We plotted the output of ''MulticlassClassificationEvaluator'' for Decision Tree and Random Forest. We can understand from the graph that baseline model 2 has got the least accuracy and Random Forest has got the most.  
523: \begin{figure}[htb]  
525: \includegraphics[width=1.0\columnwidth]{images/Figure23.png}  
527: \label{fig:Figure23}

534: A penalty matrix is defined to calculate the loss to the bank. Penalty will be applied to each misclassifications and penalty value differs for wrong classifications - 'good as bad' and 'bad as good'. As discussed in the project goals section \ref{Project goals}, approving loan for an uncreditworthy customer will end up in more financial loss for the bank and accordingly is a greater risk. Hence classifying a bad customer wrongly as good customer will have more penalty.

536: \begin{table}

539: \label{tab:table4}

550: Table \ref{tab:table4} shows the penalty matrix. For right predictions penalty is 0. If a good customer predicted as bad, the penalty is 1 and if a bad customer predicted as good, the penalty is 5. The sum of the penalty values multiplied with the respective number of misclassified customers will provide the total amount of loss / penalty.

552: Figure \ref{fig:Figure24} shows the penalty comparison of different predictive models. Baseline model 1 has more chances of predicting bad customers as good because it blindly assumes that, all the incoming customers are good. Hence it has got more penalty value. Baseline model 2 has got least chances of classifying bad customers as good because it assumes all incoming customers are bad. Hence baseline model 2 has minimum penalty.

554: \begin{figure}[htb]

556: \includegraphics[width=1.0\columnwidth]{images/Figure24.png}

558: \label{fig:Figure24}

563: Figure \ref{fig:Figure25} shows the accuracy and penalty comparison for all the 4 models. Random Forest is the most accurate model and with minimal penalty. Hence Random Forest is the best model out of all.

565: \begin{figure}[htb]

567: \includegraphics[width=1.0\columnwidth]{images/Figure25.png}

569: \label{fig:Figure25}

597: \subsection{Makefile}\label{Makefile}

605: \subsection{Project Code}\label{Project Code}

figures 25  
 tables 4  
 includegraphics 25  
 labels 36  
 refs 34  
 floats 29

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  
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 and the Customer Experience Journey

Ashley Miller  
Indiana University  
admille@iu.edu

## ABSTRACT

A customer's experience journey consists of multiple touchpoints along the way as they make choices in which companies and brands to interact with and ultimately, purchasing decisions. While the customer experience journey may differ based on product, service, audience, time, as well as a company's capabilities and strategic initiatives, the need to understand the customer transcends all industries. These touchpoints are increasingly moving to the digital space through online search, mobile interaction, social media, email, in addition to other methods that may not even be in existence as of yet. Given the number of these touchpoints across customers and the ability to track customers across multiple methods, understanding the experience of customers through the use of big data provides opportunities for companies to better enhance the customer experience journey. Real-time recommendations, personalized marketing messages, and geo-targeted advertising can all play a role in *nudging* the customer appropriately when companies are looking to drive customer interaction and behavior. We will seek to explore this customer experience journey in the digital environment and introduce relevant case studies where companies and industries have started to utilize big data and analytics to better understand and customize the customer experience journey through digital efforts.

## KEYWORDS

i523, hid329, big data, analytics, customer experience journey, consumer behavior, digital marketing

## 1 INTRODUCTION

The customer experience journey has been largely explored from a psychological and behavioral standpoint [50]. Dating back to the near 1800s, marginal and expected utility of actions were detailed by Nicholas Bernoulli, among others, to better understand how purchasing decisions were made [50]. From related work that followed in this field of studying behavioral economics, research has shown that purchasing decisions are not linear and at times, are not even rational as cognitive, emotional, and social factors can all play a role into how a customer makes a purchasing decision [50]. As described by Stoicescu, the reason why researchers started to study purchasing behavior was due to the "diversification of need" [50].

However, with diversification also comes complexity. The more choices a customer is given, the harder it can become for them to make a decision [50]. With every product choice, there also is an opportunity for interaction or *touchpoints* along this customer experience journey [34]. These series of touchpoints can occur through a variety of ways and the time frame in which they take place can also vary greatly by the product or service being offered and to which audience. In figure 1 example of a customer setting

up utilities after the purchase of a new home and the multiple touchpoints they may encounter along their journey [41].

[Figure 1 about here.]

However, not all touchpoints are created equal [34]. There are some touchpoints that every customer may have to go through to get to the next step in the process and others that will produce a more valuable action, such as a purchase [34]. There are further questions today that did not exist in years past due to the advances in technology and how that affects customer behavior [29]. These advances in technology not only could influence customer behavior but also provide companies direction on which products they should produce, where these products should be placed, what price point is most optimal and how should they properly promote a particular product to their audience [29]. Big data and analytics can provide opportunity to inform the promotional piece as companies have utilized this feature to provide personalized and relevant content along the customer journey as defined in figure 2 [50].

[Figure 2 about here.]

While traditional advertising and marketing methods have included outdoor, print, television, and radio, among others, there is a growing shift in reaching customers via the digital space [29]. As of 2016, digital advertising spend reached 72 billion dollars, a 20% increase from the year prior and now accounting for a third of all advertising revenue spent in the United States [5]. With the increased move to digital from more traditional outreach methods, the customer relationship is also being managed via digital platforms such as email, social, and mobile [29]. A customer's *digital journey* can provide opportunities for big data and analytics to better understand the touchpoints along the way as well as where a company may be able to *nudge* a customer appropriately to make a purchase decision.

The objective of this work is to provide a view into the customer experience journey as it relates to big data and analytics. This overview of existing work is to allow one to see how a company or industry may start to match their big data efforts with the purchase decision that customers make as well as the multiple touchpoints included along the way. The move towards digital outreach and marketing efforts will also be defined to ensure the reader understands what is meant throughout as it relates to outreach and personalization efforts. Rather than the analysis of a specific dataset, real-world examples will be showcased across a variety of industries to provide detail as to how big data is being used to better understand the customer and enhance the journey they go through along the way. Lastly, this work will highlight the need of matching big data with the customer experience journey, challenges with pursuing this work going forward, along with recommendations on how to overcome these challenges.

## 2 WHAT IS THE CUSTOMER EXPERIENCE JOURNEY?

The way *customer experience journey* is defined can differ by industry, product, and even by place. While past work has defined the customer experience journey as the process of purchasing a product or service, in today's landscape, it has become more than that. The Harvard Business Review would define the customer experience journey as the "sum-totality of how customers engage with your company or brand, not just in a snapshot of time, but throughout the entire arc of being a customer" [42]. Traditionally, the customer experience journey and buying process were used interchangeably where a customer moves through a decision making process. Some key areas that were highlighted in a typical customer experience journey include:

- **Need Identification:** At this stage, this is where the customer decides whether they have a particular need that they believe the product or service could fill. There are times where properly identifying the need or problem can be an area of opportunity for a company [13].
- **Awareness:** In order for customers to even engage with a product or brand, they first need to be aware that it exists. Further, the customer has to decide whether the product, service, or brand is relevant to them [13].
- **Evaluation of Alternatives:** Here, a customer starts to investigate the options available and educate themselves on the benefits and drawbacks of each. This is an area where companies seek to differentiate themselves from competitors as customers go through this stage in the progress [13]. As customers continue in their research state, they can be influenced in a variety of ways such as through advertising and marketing, word-of-mouth or reviews from others, in addition to information they obtain in other ways through their own search process [29].
- **Purchase:** After a customer has gone through their choices to the best of their satisfaction, they move to the stage where they decide to make (or not make) a purchase [29].
- **Post-Action Evaluation:** After a decision is made, one way or the other, this is place where the consumer evaluates their decision which may include key questions such as [13]:
  - Were my needs adequately met?
  - Am I satisfied with my choice?
  - If given the same circumstances, would I make the same choice again?
  - Would I recommend the choice I made to others?

While the list of questions could be endless the intent is to move customers through this purchase decision process so companies create loyal customers and advocates [29]. However, that model is evolving with the shift to a multi-prong outreach approach via digital and non-digital methods [13]. A longer customer experience journey is outlined in figure 1 as a customer can enter at any stage in the process. Pre and post purchase measures can be collected, stored, and analyzed at any point along the way as shown in figure 3 [13].

[Figure 3 about here.]

## 3 WHAT DOES DIGITAL MEAN?

With the influx of big data, analytics, and technology, there is often a rallying cry among leadership teams for an effort to be more *digital*. However, when exploring the definition of *digital*, it can greatly differ by audience, industry, and objective. Even within a singular company, alignment on what digital means can vary. Instead of trying to decide what digital *is*, it may be more beneficial to think of what digital *can do*. McKinsey highlights that digital should create value of some kind and offers various ways in which this can apply to an organization [14]. Despite varying definitions, methods, and applications of what digital *is* (and *is not*), there are commonalities in digital efforts that can be used to better understand this environment [26]:

- **Customer-Centric:** Digital efforts entail putting the customer first as they examine data and processes to enrich the customer experience [26].
- **Real-Time:** Long gone are lag times between collecting, analyzing, and processing data to decision-making. Now, data collection is always on and can be pulled at any time [26].
- **Connected:** With the volume, veracity, variety, and velocity of big data alone, the ability to have data sources and storage systems talk to one another is crucial to develop meaningful insights and inform decision making appropriately and effectively [1]. This not only has to happen from a technology standpoint but also from a company culture standpoint as well to ensure appropriate units and individuals are also talking to one another to better understand the customer experience journey overall.

It's important to note that while *digital* can mean *online*, one should not assume that these actions and behaviors only occur in the online environment. However, the collection, storing, and analyzing of these behaviors can be *digital* or *online* even though they may be reflective of what is happening in an *off-line* setting. Overall, implementing digital capabilities should improve business processes, challenge new ways of thinking, and deliver ways to enhance the customer experience journey [1].

## 4 WHAT IS DIGITAL MARKETING?

While advertising and marketing methods go as far back as the 1800s where customer lists were used to determine how individuals could be influenced via direct mailing efforts, digital marketing has only come to be with the creation of the internet [10]. The internet has created opportunity for brands to directly connect with customers and likewise, for customers to engage with brands in a myriad of ways in the digital space [17]. While varying definitions of digital marketing exist, it is often categorized as a subset of traditional marketing where the "use of digital technologies create an integrated, targeted, and measurable communication" to not only attract potential customers but engage with current ones for retention and loyalty purposes [23]. Digital marketing became even more prevalent in the 2000s as companies such as Google, Yahoo, and Facebook provided opportunity to deliver ads at an individual level based on demographic and behavioral characteristics [10]. Other data collection firms offered the ability to track users across the web space to see which pages were viewed, clicked, and time

explored to help further understand the experience of a customer across the web space [10]. With these advances in technology and understanding, the customer experience journey began to also transform along with the changes in advertising from traditional to more digital.

## 5 HOW IS BIG DATA INVOLVED IN THE CUSTOMER EXPERIENCE JOURNEY AND DIGITAL MARKETING?

As the typical customer experience journey moves away from a traditional linear process and more towards an iterative approach, there is a need to understand the pathway among customers with data [17]. As of 2016, there are now approximately 3.7 billion individual users on the internet [3]. This population size coupled with the multiple methods of interaction present an opportunity for companies to better understand potential customers in an effort to deliver the right message, to the right person, at the right time. A Gartner report states that the amount of data companies are collecting is growing by nearly 40 percent year-after-year [30]. Nearly one of out of every four state that there is a need to tie big data back to marketing-related efforts [30]. At the time of the report, it was estimated that only three percent of companies surveyed had a dedicated individual responsible for big data analytics and customer intelligence insight [30]. As the touchpoints with customers increase and also move more towards digital, so does the ability to collect and track the data from these touchpoints to better understand the customer experience journey [17].

One could argue that marketing data has been *big* for quite some time given the sheer number of people exposed to efforts typically exceeds millions [10]. However, what has changed in this space is marketing's increased use of digital technologies to reach potential customers in the digital landscape across various channels including search, display, social media, email, etc. [29]. For companies to be successful in utilizing big data to inform digital marketing efforts and to better track and enhance the customer experience journey, incorporating the necessary technologies and talent is a must along with shifting the organizational culture to making data driven decisions [23]. This process can be difficult as an individual user alone can generate "billions of data signals" and attempting to understand which ones may be tied directly to a product or brand's marketing efforts can be a daunting and overwhelming task [10]. However, there are a few well-known companies that have started the shift in tracking the customer experience journey through big data analytics and applications. We will examine key industries that have leveraged this knowledge and insight to better understand and enhance their relationship with customers.

## 6 BIG DATA IN ONLINE RETAIL

Companies like Amazon and eBay are often cited as pioneers in utilizing big data analytics considering these were companies that were *born digital* [3]. While other retail companies have made their way into the big data analytics environment, often times they are brick-and-mortar establishments in addition to offering electronic commerce (e-commerce) capabilities, such as Target, Sears, or Wal-Mart. Growth in e-commerce has taken place around the world

with nearly 1.3 billion customers in existence as of 2016 [3]. In the online space, there are multiple opportunities for these companies to track the customer experience from number of visits, keywords used in search, orders and products placed, frequency of purchases, in addition to when items in their virtual cart are abandoned or even when items are returned or complaints are filed [3]. Amazon and eBay, among others, have utilized big data analytics to their advantage to create recommendations for their customers, develop predictive models, and also offer real-time changes in the customer's experience journey.

### 6.1 How Does Amazon Use Big Data?

In 1995, Amazon started its life in the online space as an electronic bookstore [25]. While early sales were still impressive totaling nearly \$20,000 a week, during their popular campaign of *Prime Day* in 2017, analysts estimated that sales for that one day alone to be around \$500 million [47]. As of July 2017, Amazon has approximately 300 million users with nearly eight out of ten that make a purchase at least once a month [47]. The amount of information tracked per user continues to increase at an exponential rate as Amazon's growth has moved beyond their days of an online bookstore and into the realm of an *everything store*, including acquiring other large companies such as Whole Foods [40]. With this, there are a number of ways that Amazon uses big data to enhance the customer experience:

- **Personalized Recommendations and Predictive Modeling:** As customers are exploring products on Amazon, they are often shown other suggested products either based on their own past purchase behavior or by what others who purchase similar items also bought [3]. These recommendations are shown in real-time, often on the same web pages as customers are exploring other products [3]. It is estimated that based on this ability alone, utilizing both structured and unstructured data, that 35% of all sales are attributed to the recommendation algorithm, which would show that their predictive efforts are meeting the needs of their customers [3].
- **Efficient Delivery:** Even though Amazon utilizes and houses large amounts of data about its customers to offer personalized recommendations and offers, the company also has to maintain a tremendous amount of data about its own operations to inform logistics and supply chain management to meet customer expectations [3]. A key selling message of Amazon is delivering product in two business days, even in some cases offering same day delivery on certain items if they are ordered within a certain timeframe CITE[21]. With the increase number of products, suppliers, and customers, Amazon is still able to rely on big data analysis to maintain a consistent experience that doesn't compromise delivery for the customer which in turn leads to a better customer experience overall [3].

### 6.2 How Does eBay Use Big Data?

eBay is a website that also started in 1995 which offered a unique opportunity to bring together buyers and sellers in the online space [15]. Sellers could place their items online where buyers could

potentially place bids, similar to if they participated in a live auction, where the item may go to the highest bidder. This allowed for others around the globe to connect and purchase items directly from another person while paying for the item through online methods. It is estimated today that there are nearly 180 million buyers and sellers and nearly 250 million search inquiries made per day on eBay [31]. Like Amazon, eBay seeks to understand and tailor the online experience for customers through the use of big data in a multitude of ways as eBay itself states “understanding the customer is key” [44]. Various methods utilized to better understand and tailor the customer experience journey include:

- **Web Page Metrics to Inform Layout:** It is estimated that among eBay customers, there is “100 million hours of interactions collected per month” [44]. Through an extensive number of experiments and A/B testing, eBay is able to optimize the web experience for customers. From their big data analytics, they can find preferred layouts of web pages which can customize anything from navigation feature to the size of photos displayed on the screen [3].
- **Ease of Finding Items:** Buyers and sellers alike utilize the search feature provided on the eBay website to find necessary items or to compare price points of like items when deciding which price point to utilize [31]. Behavior patterns of customers have been used to inform how to best optimize the search feature in an effort to get customers to the necessary items more quickly which in turn will, hopefully, produce a sale [31]. While in the past, the search algorithm would have taken words and terms in a more literal sense, though optimization eBay has been able to make the search algorithm more intuitive which has lead to more sales [31]. Such examples show that originally when customers would shorten words used in the search feature, they may not find what they need. However, after analysis of customer inputs and changes in the search algorithm, this ability was taken into account so customers could still find the necessary product without changing their behavior.

## 7 BIG DATA IN THE FINANCE INDUSTRY

There are a number of products and services available in the finance industry ranging from personal and business loans, to stocks, retirement accounts, and credit cards. Companies such as JP Chase Morgan, American Express, and Bank of America are capitalizing on big data use to inform their offerings and also better understand their customer base [52]. These companies are monitoring the customer experience journey through all touchpoints which could include web visits, phone calls, and even in-person interactions [24]. This information can also be used to detect fraud on certain accounts when activity occurs that may not be typical of their customer [52]. These algorithms and techniques can in turn ensure customers are protected which help with customer retention. Conversely, those in the finance industry may also be able to utilize big data and mining techniques to determine if they are about to lose a customer. One such company that utilized these methods to their advantage is American Express, which accounts for nearly

a quarter of all credit cards transactions and totals more than \$1 trillion annually in customer purchases [35].

### 7.1 How Does American Express Use Big Data?

In 2010, American Express invested in big data technologies and resources, including Hadoop, to increase capabilities to detect fraud, provide recommendations to current customers, predict who may close their accounts, as well as acquire new customers [52]. These methods are used to assist in efforts to maximize the customer experience journey through the use of:

- **Fraud Detection:** To minimize loss, fraud alerts have to happen quickly. To achieve this, American Express implemented machine learning algorithm techniques [32]. Data points included in the model consisted of information about the merchant where the purchase occurred, purchase details such as items bought and price, and even customer information [32]. By analyzing patterns in real-time, American Express was able to flag possible fraudulent activity in a matter of milliseconds which then allows the company and customers more time to prevent further loss with the increased capability [32]. American Express was able to identify an additional \$2 billion in fraudulent activity that that they were not able to identify before and therefore protect their customers and ensure a more positive customer experience as a result [32].
- **Personalized Recommendations:** Along with protecting their customers, American Express also seeks to understand how to better engage their customers through personalized recommendations. One such example is based on analyzing customers past transactions along with geographic location information to push specific recommendations to customers in real-time [52]. Through the use of big data analytics, the company can send recommendations on similar restaurants in the area if they see from a customer’s transactions they frequent a certain genre or area [52]. These recommendations also work on behalf of the merchants who accept American Express as the company can provide information on purchases in the area which merchants can use to create offers to entice customers to purchase products and services by using their American Express card at their particular store [16].
- **Churn Prediction Models:** American Express also uses its vast amounts of data to see if they can predict whether a customer will close their account [32]. By incorporating machine learning models, they can better understand the customer experience and appropriately jump in at different points along the journey in an effort to deter customers from closing their accounts. Through analysis of past transactions as well as nearly 100 other variables incorporated to understand customers, the company estimated that for one model, they were able to identify nearly one out of every four accounts they believed would have closed in the near future [32]. With this information, tailored marketing and messaging could be implemented to help with retention rates.

- **Acquiring New Customers:** Despite the large base of customers, merchants, and transactions, there is always a need for businesses to grow to increase revenue and capabilities for the future. One way to achieve this is through digital marketing efforts targeted at those who may be potential customers for American Express. Through their efforts, American Express was able to grow their customer base by nearly 40% through online marketing efforts [32]. With these more targeted and cost-effective measures, American Express was able to efficiently acquire new customers as compared to more traditional marketing efforts of the past, such as direct mail [32]. These optimizations further enhance the customer experience journey by delivering them a message at the right time through the right medium.

## 7.2 How Does Bank of America Use Big Data?

While big data, analytics, and predictive models can be used to better understand how to reach out, retain, and attract customers, these same techniques can be applied when determining how to optimize the customer experience journey from an internal perspective. Considering the journey can take place across a series of touchpoints, Bank of America was one major bank, among others, that utilized big data to better understand how to better serve their nearly 50 million customers [12]. One method included:

- **Customer Segmentation:** Through the use of big data, Bank of America acknowledged that their customer base could be divided into segments and therefore their behavior and needs differed [12]. By analyzing online correspondence, calls from a call center, and even visits to area branches, appropriate offers could be tailored to the customer [12]. Utilizing data points provided in the online space along with the ones that occurred elsewhere, a new program was developed by Bank of America [12]. With this new program and customized offerings, customers were more highly engaged with by Bank of America which increased customer satisfaction and experience as a result [12].

## 8 BIG DATA IN THE HEALTHCARE INDUSTRY

Rising patient volumes, increasing aging population, and mounting costs have all contributed to the growth, importance, and complexity of the healthcare industry [37]. As of 2016, it is estimated that nearly \$4.1 trillion will be spent on healthcare costs in the United States alone [37]. Nearly 290 million people in the United States have some form of insurance or healthcare coverage but that also leaves nearly 28 million who are uninsured [19]. A typical customer can interact with a number of stakeholders throughout their healthcare journey ranging anywhere from their initial doctor's visit, to filling a prescription at the pharmacy, to paying a bill to their insurance provider. One may then ask based on these interactions: how does the online space play into the healthcare industry at all?

With the move to electronic medical records (EMR), the ability to now aggregate years of information on an individual, as well as an entire population, becomes more of a reality [20]. Even though

the healthcare industry has lagged behind other industries regarding their collection and use of big data, they are one of the most important as it relates to utilizing their information to create a better experience for customers as it pertains to their health [20]. The ability to link this data across various stakeholders is also critical in understanding the full journey of a customer (or patient) to ensure effective treatment decisions are made. Health Information Exchanges (HIEs) allow for this opportunity and the HIE has information on more than 10 million patients, over a span of nearly 80 connected hospitals, and approximately 18,000 physicians have access [20]. Big data in the realm of healthcare provides tremendous opportunity to create value for customers and healthcare professionals alike. One such software company explored this use of connected data sources to better inform healthcare providers with practice-based evidence in an effort to tailor care for an individual patient [6].

## 8.1 How Does Apixio Use Big Data?

As others have stated about healthcare related data and reporting, "the problem in healthcare is not lack of data, but the unstructured nature of its data" [33]. Apixio, a cognitive computing firm based in California, wanted to take on the challenge of making unstructured healthcare related data available and easier to use in order to better aid decision making in patient treatment [33]. Their work involved taking clinical charts of patients and combining them with notes from physicians, test results, and even hospital stays to develop a more complete picture about an individual [33]. From there, Apixio was able to provide benefits based on this big data process:

- **Patient Model Development:** Data at an individual level was used to develop patient models from a series of text processing and coded healthcare data [33]. By creating a profile per individual, like individuals could then be grouped together which in turn helped to inform what treatments or procedures would work best in those individuals who fit a certain criteria [33]. Considering this information is derived from actual practice of medicine, it can better inform clinical care and also ensure that patients are set up for best optimal outcomes if treatment decisions are made based on big data collection and analysis [33].

- **Healthcare Cost Savings:** Cost of healthcare continues to be a growing concern for both customers and other key players such as healthcare professionals and insurance companies [37]. With the move to EMR and big data analytics, it is estimated that anywhere between \$300 and \$450 billion dollars can be saved in healthcare costs [37]. With the use of big data technology and methods, Apixio developed a system that could read and code patient chart information [33]. Typically, this method of coding would have been performed manually by a person or set of individuals, and with that comes a laborious and expensive process [33]. Apixio's capabilities were also found to be more accurate resulting in 20% improvement in accuracy which in turn lead to better decision making among healthcare providers [33]. The also helped individual customer to ensure they were getting billed appropriately for the right treatment or procedure as well as for the insurance

company who may be providing coverage [33]. These techniques then allow for an improved customer experience journey if costs can be mitigated through the use of big data initiatives that allow for better efficiency and accuracy.

## 9 BIG DATA IN THE ENTERTAINMENT INDUSTRY

The entertainment industry includes a wide array of forms including newspapers, movies, books, television programs, and radio [22]. As of 2016 it is estimated that this industry is worth approximately 1.8 trillion dollars in the United States alone [49]. Streaming video services such as Netflix and Hulu have entered the market in recent years and provide a further opportunity to deliver content directly to customers. As of 2016, video streaming services are the second largest category for home entertainment with customers in the United States spending \$6.2 billion [4]. The wealth of data collected from these streaming services include but are not limited to the type of content watched, when content is watched and on which type of device, as well as how often it takes for customers to make a selection down to an individual user level [8]. Netflix is one of the many video streaming leaders and has made big data and analytics a foundation to their business strategy and outreach initiatives [28].

### 9.1 How Does Netflix Use Big Data?

While Netflix once started out as a mail-subscription video rental service, the business model has shifted to provide content entirely online and caters to nearly 60 million subscribers in over 50 countries [28]. Netflix's competitive advantage in the market place stems from their ability to use big data as they estimate that they process over 10 petabytes of data a day which includes more than 400 billion new events [28]. Utilizing programs and data scientists, Netflix began to seek out additional opportunities to understand customer preferences and to also optimize the experience journey through a variety of different methods:

- **Personalized Recommendations:** Netflix not only analyzes what a particular person may watch but also what others who *look like* that user may enjoy based on data such as age, gender, or even zip code [28]. With the sophistication of the recommendation algorithm, viewers spend an average of 17.8 minutes browsing through the selections before picking a program to watch [28]. Spending more time increases the level of engagement with users and also extends the lifetime value of the customer in an effort to help with retention [8]. By delivering relevant content, Netflix estimates they save more than \$1 billion per year by their efforts in keeping customers happy [8].
- **User Choice:** In addition to providing the right recommendations, ensuring that the image or artwork for films is appropriate to the user also aids in choice [8]. Netflix engages in A/B testing of program thumbnails images and also seeks out feedback from users on which images they prefer [8]. From this process, Netflix was able to increase video viewing between 20-30% when utilizing the right images and listening to customers' preferences [8].
- **Customized Content:** Analyzing what audiences enjoy watching can provide insight as Netflix sought to create

their own content [28]. One common cited example includes the development of *House of Cards* as an original Netflix series that was created with big data information [28]. Netflix found that the original series from the British Broadcasting Corporation (BBC) did well with audiences and that Kevin Spacey movies were also popular [28]. Further using customer data, Netflix understood that customers *binge-watched* seasons of shows and therefore releasing an entire season at a time would best meet the needs of their customers versus one episode at a time [28]. The year the *House of Cards* series premiered, subscribers grew from 27.1 to 33.4 million and the show received countless Emmy and Golden Globe nominations and awards [28]. By utilizing big data, Netflix was able to create and deliver content that customers wanted and also help their bottom line [28].

## 10 BIG DATA IN THE GAMING INDUSTRY

In addition to the entertainment economy, the gaming sector also is substantial in size and revenue. In 2016, the commercial gaming industry grew to \$38.7 billion across 24 states and nearly 600 casinos [43]. Las Vegas, a leader and popular gaming destination had a record year of visitors at nearly 43 million [43]. With increased competition among entertainment resorts and casinos in Las Vegas, as well as other parts of the United States, the need to create an optimal customer experience is crucial to attract customers and also keep them engaged. Metro-Goldwyn-Mayer (MGM) Resorts International and Caesars Entertainment are two conglomerates that have capitalized on big data use to better tailor the customer experience journey.

### 10.1 How Does Caesars Entertainment Use Big Data?

Caesars has described their customer relationship optimization process as utilizing a "data-driven and closed loop approached to deliver a personalized experience" [51]. A few ways they have implemented this include:

- **Creating Customer Loyalty:** Demographic, gameplay, and other transactional data is kept on each guest to create a detailed profile [51]. Employees then across the establishment can utilize this data to personalize offerings and incentives to customers, anywhere from how he or she is greeted by staff to whether complementary services should be offered to improve the customer experience [51]. This type of treatment isn't just limited to big spenders at the casino but translates across all customers in an effort to create loyalty across multiple segments [51].
- **Efficiencies Through Mobile Application:** Caesars also offers guest the ability to utilize their mobile device to conduct tasks such as checking into a property or even ordering a drink from the bar to avoid long lines [51]. Incentives can also be pushed directly to customers based on their location and preferences such as tickets for shows or dining options in the area [51]. Considering most guests carry their phone in their pocket, engaging with them on

the casino floor can create a better customer experience to give them what they need, when they need it [51].

- **Customized Experiences:** The vast amounts of data collected on customer behavioral patterns in terms of which machines are played, when, where, and by whom can provide insight into how to best tailor offerings [46]. For example, it was observed that an elderly population visits the casino at a certain time of day and therefore with the influx of that audience, casinos are able to adjust game offerings in real-time offering enlarged text for better viewing among the visually impaired in that age group as well as bet levels for certain games [46]. By analyzing heat maps of popular games and parts of the casino, it also allows for companies to staff appropriately to ensure customer needs are being met based on predicted demand [46]. These real-time changes enhance the customer experience journey by tailoring offerings to specific customer segments.

## 10.2 How Does MGM use Big Data?

Similar to other casinos and resorts, MGM has utilized past customer data in an effort to better predict future behavior [36]. However, they have also utilized this data to create personalized marketing offers to entice frequent (and non-frequent) visitors back into the experience [36]. Though sophisticated modeling efforts, MGM is able to tailor marketing efforts to include a variety of different incentive types and levels. The final result of this process created 120 models of customer behavior with approximately 180 variables in each as well as 20,000 parameters across all which showcased an increase in revenue at a lower cost [36]. These models were used to inform marketing efforts across a variety of areas, including but not limited to [36]:

- **Hotel Room Rates:** Attributes such as room type, discount, number of times, etc., all play a role in which aspect will draw a customer back into the establishment [36].
- **Entertainment Add-Ons:** Type of entertainment offered, ticket price, or even facility features were all used as inputs in the model [36].
- **Other Offers:** Air packages, limo rides, resort credits, and many others were also used as way to determine which customers would respond to which offers [36].

## 11 BIG DATA IN THE TRANSPORTATION INDUSTRY

Whether by plane, train, car, or other means, today's American customer relies on some sort of transportation to get them to varying destinations whether it be work, school, or even vacation. An average person spends 20% more time commuting today than they did 30 years ago [7]. With this come questions to the transportation industry as to where they should expand highways, add public transit, or open additional hubs or destinations for travel. Big data can be one avenue in exploring and answering these questions as well as create a more enjoyable experience for the customer if they can spend less of their life commuting. The introduction of the ride sharing mobile application of Uber also arose based on customer needs and preferences and through the use of big data is thriving as alternative transportation option [9]. Large airline carriers have

made use of big data as they seek to understand buyer behavior so they can effectively plan flights and other amenities to meet customer needs [39].

### 11.1 How Do Airlines Use Big Data?

In just one day's time, it is estimated that there are nearly 42,000 commercial flights and 2.5 million passengers [2]. From purchasing a ticket, to taking a flight, and (hopefully) receiving their checked baggage at their final destination, airlines collect a wealth of information on their customers throughout their flying experience [39]. When looking at key attributes that are analyzed and down to an individual level, airlines collect information about purchase history, arrival, departure cities, and dates, in-flight food choices, connecting cities, travel companions, as well as miles and credit card points earned and used [18]. While airlines have succeeded in collecting this data, using it to better enhance the customer experience journey is still a work in progress [39]. Those who work in the travel software environment and frequently provide products and services to those in the airline community to better understand their data even state they have "not seen a single major airline with an integrated big data business solution" [39]. With that in mind, highlights from major airline players are explored even though full development of utilizing big data may still be on-going in this industry.

### 11.2 How Does Southwest Airlines Use Big Data?

One way that Southwest Airlines is utilizing big data is by trying to identify new opportunities for revenue [39]. By analyzing customer behavior online, Southwest is able to support their relationship with customers by offering the best rates in real-time [39]. They are also able to look at searches for destination pairs and make determinations on whether certain flights should be added to keep their customer base loyal and ultimately satisfied by getting to where they need to be, when they need to be there [18]. Not only is Southwest looking to meet the needs of customers as they make a flight choice, but they also seek to comprehend customer interaction at other points in the purchase process [18]. By utilizing a speech analytics tool, the company can better understand recorded conversations that take place with representatives as well as social media chatter [18]. These real-time insights can then inform customer service representatives as they interact with customers and guide them to deliver the optimal solution in various situations [18]. In addition to optimizing the customer experience from a satisfaction standpoint, Southwest airlines has also partnered with NASA on potential safety initiatives where machine learning algorithms can be used to spot potential abnormalities [18]. These efforts enhance the customer experience journey by not only looking out for safety of individuals but by also meeting their needs based on behavioral data.

### 11.3 How Does Delta Airlines Use Big Data?

In a quest for customer loyalty, Delta Airlines has made an intentional effort in investing in their baggage tracking system to better meet customer needs [45]. With this \$100 million dollar initiative, it not only gives airport operation teams the opportunity to identify

trends in mishandled luggage situations but also real-time information to baggage handlers when transferring or sorting through bags [45]. Similar information is also shared with travelers so they can track the progress of their bags down to the minute [45]. With approximately 130 million bags checked in a given year on Delta Airlines, there is a common concern among customers on whether their bag will arrive at their final destination [39]. Giving customers a piece of mind allows for a more beneficial customer experience and also increases satisfaction and loyalty. The luggage tracking app has been downloaded 11 million times and has reduced the rate of mishandled luggage by nearly 71% since 2007, which is better than any other airline [45].

#### 11.4 How Does Uber Use Big Data?

Founded in 2009, Uber started as a technology company and created a mobile application that connected those seeking transportation with those who were drivers [9]. Now, with nearly 8 million users who have connected with over 160,000 drivers, nearly half of the United States population has access to Uber in their city [27]. The only opportunity to connect riders and drivers is through the mobile app consolidating data collection and tracking from the start; however, the sheer volume and real-time application of data use to inform pricing and availability still presents on-going challenges [9]. Demographics, frequency of trips, destinations, price, as well as sessions that do not end with a purchase are all recorded from the application [9]. Several ways in which Uber utilizes big data to meet customer needs includes:

- **Matching Supply and Demand:** By analyzing travel transactions, Uber can appropriately plan for busy nights so customer travel needs are met [9]. Customers are also able to give feedback about their ride experience and rate drivers [6]. With this capability, the company can inspire trust and improve satisfaction if they find that certain drivers are not meeting expectations [6]. UberPool is also a new feature that has been added that allows for carpooling of customers where real-time analytics search for other customers in the area by geography [6]. This can therefore improve the customer experience for those who want to share a ride and split the cost appropriately [6].
- **Dynamic Pricing Model:** Uber is also able to adjust pricing models accordingly based on time of day [9]. Fair estimates are also able to be given in real-time which allow the customer to adjust their travel plans if needed and also pick the type of transportation, such as a sedan or sports-utility-vehicle (SUV) [9]. However, there are times that Uber uses these models to the company's advantage and offer *surge* pricing in the events of heavy demand or traffic [9]. All financial transactions take place via the application with no exchange of cash. Pre-set and transparent pricing structures allow customers to select what fits their needs, even if they find their choice is to not take a ride at a particular time. Having the necessary and accurate information provided at time or purchase makes for a more enjoyable customer experience journey [6].

#### 12 WHY IS USING BIG DATA TO OPTIMIZE CUSTOMER EXPERIENCE JOURNEY IMPORTANT?

These examples are a select few to showcase how companies can better understand and further the customer experience journey by leveraging big data. The average customer is presented with more choices today than ever before [34]. With this, companies today have to be more strategic to get the attention, time, and loyalty of customers to remain in the marketplace. Doing so can provide many advantages to both companies and customers as they utilize big data to better understand the customer experience. As Rawson et al state: "companies that excel in delivering journeys tend to win in the market" [41]. Trends presented showcase how big data can provide big benefit:

- **Retention of Customers:** It is estimated that "acquiring a new customer can be between five and 25 times more expensive than retaining an existing one" [21]. Utilizing big data to predict when customers may close accounts can help to inform company efforts and ultimately prevent potential revenue loss if they can keep existing customers. American Express showed that by using big data and predictive analytics, the company could identify these customers sooner versus wait until the customer is already lost.
- **Personalized Outreach:** Tailored communication messaging, and recommendations can give customers a better experience in getting what they need from companies but also benefit the company's bottom line as well. As Netflix and Amazon have showcased, providing recommendations to customers increases engagement and purchase behavior.
- **Company Process Efficiencies:** Utilizing big data to understand customer behavior can help companies determine whether changes or improvements need to be made in how they deliver products and services to customers. As the Delta example showed, tracking baggage was not only a concern to customers but by doing so, the company improved their mishandled luggage rates. These efficiencies not only create satisfied, and possibly loyal, customers but also ensure that companies are spending their resources effectively by not making costly and time-consuming errors.

#### 13 WHAT CHALLENGES EXIST IN UTILIZING BIG DATA FOR THE CUSTOMER EXPERIENCE JOURNEY?

While big data use is going to be a crucial component going forward in understanding the customer experience journey, companies do face challenges in making this a reality, specifically:

- **Data Ownership** Customer data can live in a variety of places within one organization. The departments in which this data lives can be in silos with multiple departments not talking to one another or not willing to share what they feel their department owns [48].
- **Company Culture** Cooperation across the organization can be a significant barrier in truly understanding the full customer experience. [48].

- **Lack of Strategy** Without a clear strategy, it can also create issue in trying to determine how to best interpret and apply the findings in the data. This can lead to gaps in the organization if it is unclear what the ultimate goal is and which parties play a role [48].
- **Resources and Skills** The technical aspects of understanding the customer experience journey can be a barrier as well. Having the right technology, people, and time in place to understand the full customer journey can also be a challenge for companies. [11]
- **Consumer Behavior Volatility** Not all decisions made by customers are rational ones and there can be a variety of factors in play that big data can not track [50]. As further detailed in other work “people do not behave like robots,” so even when all the variables are optimized, outside forces beyond the control of a company could influence choice along with a customer’s own emotions which big data doesn’t always include [42].

Since within one company there can be different systems, different processes, and a variety of people employed with different skillsets, trying to address all of these challenges can be overwhelming. However, with challenges come opportunities and areas in which companies can focus on as they strive to have data that is connected, customer-centric, and available to look at in real-time [48].

## 14 HOW CAN A COMPANY OVERCOME THESE CHALLENGES?

As companies seek to better understand their customer base through the use of big data and analytics, from the research performed, there are some steps that companies can take as they further explore opportunities to optimize the customer experience journey. Some key areas and questions to consider include:

- **Seek to Understand Your Customer:** Big data and analytics can be a valuable starting point in understanding the pathway to purchase among your customers as well as which areas where you may be losing customers in the process. However, big data should be used in conjunction with small data as companies seek to understand the *why* behind customer behavior. Gathering feedback from customers is essential in the process in optimizing their journey.
- **Set Clear Objectives and Roles:** Given that earlier research highlighted that a very small percentage of companies have a dedicated person for customer analytics, first establishing a dedicated person or team could help in developing a better understanding of the data involved in tracking the customer experience journey [38]. This person or team of people can provide guidance to others within an organization by being a central source of knowledge about the customer. A key part of research is also setting clear questions and objectives at the beginning. Which data points are truly a part of the customer experience journey? What connections do we need to establish in order to move our strategy forward? How will we measure the return on our efforts?

- **Make the Necessary Investment:** As other companies in this research highlighted, big data skillsets are necessary in understanding the customer experience journey which may mean a company may need to add data scientists, analysts, or other like positions within an organization. Additional technologies and tools may also be needed such as Hadoop or languages such as R or Python in an effort to process big data to derive insight.
- **Test, Assess, and Optimize:** As companies look to establish dedicated resources, time, and people in the process of understanding the customer experience journey, there must also be acknowledgement that this process is iterative. There could be efforts that are not fruitful or plainly, do not work. However, as other companies have shown, the ability to test can provide this insight and allow the opportunity for a company to change course if needed.

While there are likely other areas to consider, this initial outline described can provide companies and those within an opportunity to start understanding the customer experience journey from a big data and analytics perspective. A company has to also prioritize these different efforts accordingly as it may not be possible to implement these changes at once. A company must also consider what their own success will look like over time as progress is made.

## 15 CONCLUSION

The customer experience journey will continue to evolve as new technologies are developed that can influence the multitude of touchpoints one experiences along the way as they make purchasing decisions. While big data and analytics can provide a picture as to what customers are doing, leveraging learnings from this work to better understand the customer experience journey will be key as competition in the marketplace continues to increase across a variety of industries. These examples show that by having the right tools, skillset, and objectives in place that utilizing big data to better meet the needs of customers can be successful. While the undertaking of this endeavor may not be quick or necessarily easy, it can provide great benefit to both companies and customers to deliver relevant products and services with the customer experience in mind. Even though big data is a means of tracking the customer experience, big data is also changing the customer experience through digital marketing and outreach efforts in a way to effectively and efficiently engage and connect with customers. With this approach, the ability to deliver the right message, to the right person, at exactly the right time in the customer experience journey can provide tremendous opportunity for companies but also benefit the customers to have a more fulfilling experience journey with a company or brand.

## ACKNOWLEDGMENTS

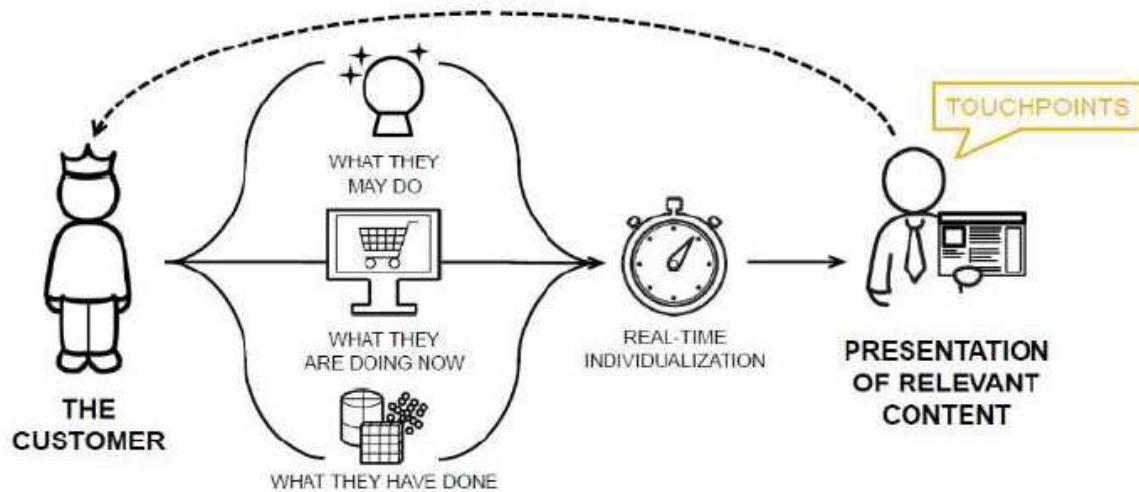
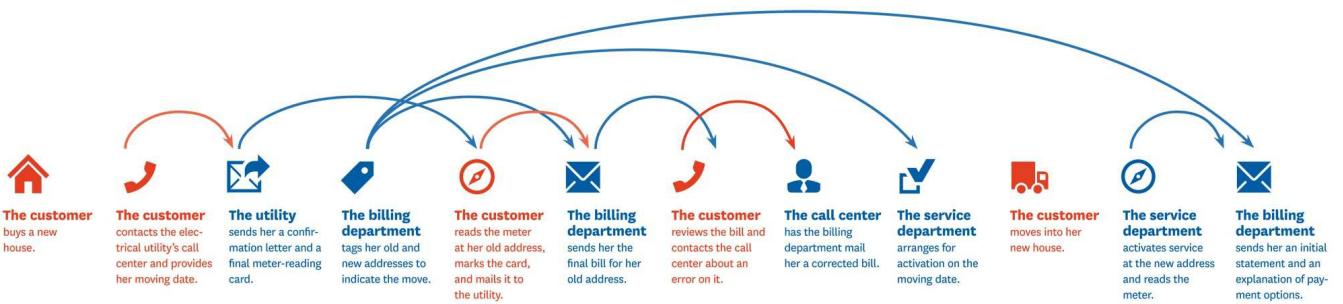
The author would like to thank Dr. Gregor von Laszewski and the teaching assistants for their support and guidance in writing this paper in addition to the resources provided by the School of Informatics, Computing, and Engineering at Indiana University in Bloomington.

## REFERENCES

- [1] Accenture. 2013. What it Means to be Digital. (2013). <https://www.accenture.com>
- [2] Federal Aviation Administration. 2017. Air Traffic by the Numbers. (2017). [https://www.faa.gov/air\\_traffic/by\\_the\\_numbers](https://www.faa.gov/air_traffic/by_the_numbers)
- [3] Shahria Akter and Samuel Fosso Wamba. 2016. Big Data Analytics in E-Commerce: A Systematic Review and Agenda for Future Research. (2016).
- [4] Claire Atkinson. 2016. Video Streaming Services Saw Giant Leap in 2016. (2016). <https://nypost.com>
- [5] Pew Research Center. 2017. Digital News Fact Sheet. (2017). <http://www.journalism.org/fact-sheet/digital-news/>
- [6] Data Science Central. 2017. The Amazing Ways Uber is Using Big Data. (2017). <https://www.datasciencecentral.com>
- [7] Tor Clifford. 2017. Five Urban Transportation Challenges that Big Data Can Help You Solve. (2017).
- [8] Jonathan Cohen. 2017. Netflix's Use of Big Data: Lessons for Brand Marketers. (2017).
- [9] Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. 2016. Using Big Data to Estimate Consumer Surplus: The Case of Uber. (2016).
- [10] Nick Couldry and Joseph Turow. 2014. Advertising, Big Data, and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy. *International Journal of Communication* 8, 0 (2014), 1710–1726.
- [11] David Court. 2015. Getting Big Impact from Big Data. (2015), 8 pages.
- [12] Thomas H. Davenport and Jill Dyche. 2013. Big Data in Big Companies. (2013).
- [13] Gary DeAsi. 2017. Why the Customer Journey is Your New Marketing Funnel. (2017).
- [14] Karel Dorner and David Edelman. 2015. What 'Digital' Really Means. (2015), 5 pages.
- [15] Ebay. 2017. Our History - Ebay. (2017). [www.ebay.com](http://www.ebay.com)
- [16] The Economist. 2016. The Economist. (2016).
- [17] David C. Edelman. 2010. Branding in the Digital Age. (2010), 8 pages.
- [18] Exastax. 2017. How Airlines are Using Big Data. (2017). <https://exastax.com>
- [19] The Henry J. Kaiser Family Foundation. 2016. Health Insurance Coverage of the Total Population. (2016). <http://www.kff.org>
- [20] Peter Froves, Basel Kayyali, David Knott, and Steven Van Kuiken. 2013. The 'Big Data' Revolution in Healthcare. (2013), 23 pages.
- [21] Amy Gallo. 2014. The Value of Keeping the Right Customers. (2014).
- [22] Brian Griffith. 2017. Playing to Win with Analytics. (2017).
- [23] Ketty Grishikashvili, S. Dibb, and M. Meadows. 2014. Investigation into Big Data Impact on Digital Marketing. (2014), 26-37 pages.
- [24] Tom Groenfeldt. 2012. Banks Use Big Data To Understand Customers Across Channels. (2012). <https://www.forbes.com>
- [25] Avery Hartmans. 2017. 15 Fascinating Facts You Probably Didn't Know About Amazon. (2017). [www.businessinsider.com](http://www.businessinsider.com)
- [26] Reda Hmeid. 2017. What Does "Being Digital" Actually Mean? (2017). <https://www.infoq.com>
- [27] Statistic Brain Research Institute. 2017. Uber Company Statistics. (2017). [www.statisticbrain.com](http://www.statisticbrain.com)
- [28] Tricia Jenkins. 2016. Netflix's Geek Chic: How One Company Leveraged its Big Data to Change the Entertainment Industry. *Jump Cut: A Review of Contemporary Media* 7, 1 (2016), 1–17. Issue 57.
- [29] P.K. Kannan and Hongshuang Li. 2017. Digital Marketing: A Framework, Review, and Research Agenda. *International Journal of Research in Marketing* 34 (2017), 22–45. Issue 1.
- [30] Kelly Liyakasa. 2013. Big Data and Customer Experience Begin to Converge. (2013). [www.destinationCRM.com](http://www.destinationCRM.com)
- [31] Spandas Lui. 2012. How eBay Uses Big Data to Make You Buy More. (2012). [www.zdnet.com](http://www.zdnet.com)
- [32] Charu Mangani. 2017. American Express: Using Data Analytics to Redefine Traditional Banking. (2017). <https://digit.hbs.org>
- [33] Bernard Marr. 2016. *Big Data in Practice*. Wiley, Corporate Headquarters 111 River Street Hoboken, NJ 07030-5774.
- [34] Christopher Meyer. 2007. Understanding Customer Experience. (2007).
- [35] Timothy Pickett Morgan. 2014. Why Hadoop is the New Backbone of American Express. (2014). [www.enterprisotech.com](http://www.enterprisotech.com)
- [36] Harikesh S. Nair, Sanjog Misra, William J. Hornbuckle IV, Ranjan Mishra, and Anand Acharya. 2016. Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation. (2016).
- [37] Raghunath Nambiar, Ruchie Bhardwaj, Adhiraj Sethi, and Rajesh Vargheese. 2013. A Look at Challenges and Opportunities of Big Data Analytics in Healthcare. In *2013 IEEE International Conference on Big Data*. 2013 IEEE Conference on Big Data, Silicon Valley, CA, USA, 17–22. <https://doi.org/10.1109/BigData.2013.6691753>
- [38] Wes Nichols. 2013. Advertising Analytics 2.0. (2013).
- [39] Katherine Noyes. 2014. For the Airline Industry, Big Data is Cleared for Take-Off. (2014). [www.fortune.com](http://www.fortune.com)
- [40] Greg Petro. 2017. Amazon's Acquisition of Whole Foods is About Two Things: Data and Product. (2017). [www.forbes.com](http://www.forbes.com)
- [41] Alex Rawson, Ewan Duncan, and Conor Jones. 2013. The Truth About Customer Experience. (2013), 10 pages.
- [42] Adam Richardson. 2010. Understanding Customer Experience. (2010).
- [43] Rubinrown. 2017. Gaming Statistics. (2017).
- [44] Cliff Saran. 2014. How Big Data is Powering Success for eBay's Customer Journey. (2014). [www.computerweekly.com](http://www.computerweekly.com)
- [45] Harvard Business School. 2015. Big Data Takes Flight at Delta Air Lines. (2015). <https://digit.hbs.org>
- [46] Natasha Dow Schull. 2012. The Touch-Point Collective: Crowd Contouring on the Casino Floor. (2012).
- [47] Craig Smith. 2017. 120 Amazing Amazon Statistics and Facts. (2017). [www.expandedramblings.com](http://www.expandedramblings.com)
- [48] Jeffrey Spiess, Yves T'Joens, Raluca Dragnea, Peter Spencer, and Laurent Philippart. 2014. Using Big Data to Improve Customer Experience and Business Performance. *Bell Labs Technical Journal* 18, 4 (2014), 3–17.
- [49] Statista. 2017. Value of the Global Entertainment ad Media Market from 2011 to 2021. (2017). <https://www.statista.com>
- [50] Christina Stoicescu. 2015. Big Data, The Perfect Instrument to Study Today's Consumer Behavior. *Database Systems Journal* 6, 3 (2015), 28–41.
- [51] Michael Welch and George Westerman. 2012. Caesars Entertainment: Digitally Personalizing the Customer Experience. (2012).
- [52] Alex Woodie. 2016. How Credit Card Companies are Evolving with Big Data. (2016). [www.datanami.com](http://www.datanami.com)

#### LIST OF FIGURES

1	Customer Experience Journey Touchpoints	12
2	Customer Experience Journey	12
3	Digital Marketing Customer Stages Model	13



## New Digital Marketing Hourglass: Customer Journey Stages Model

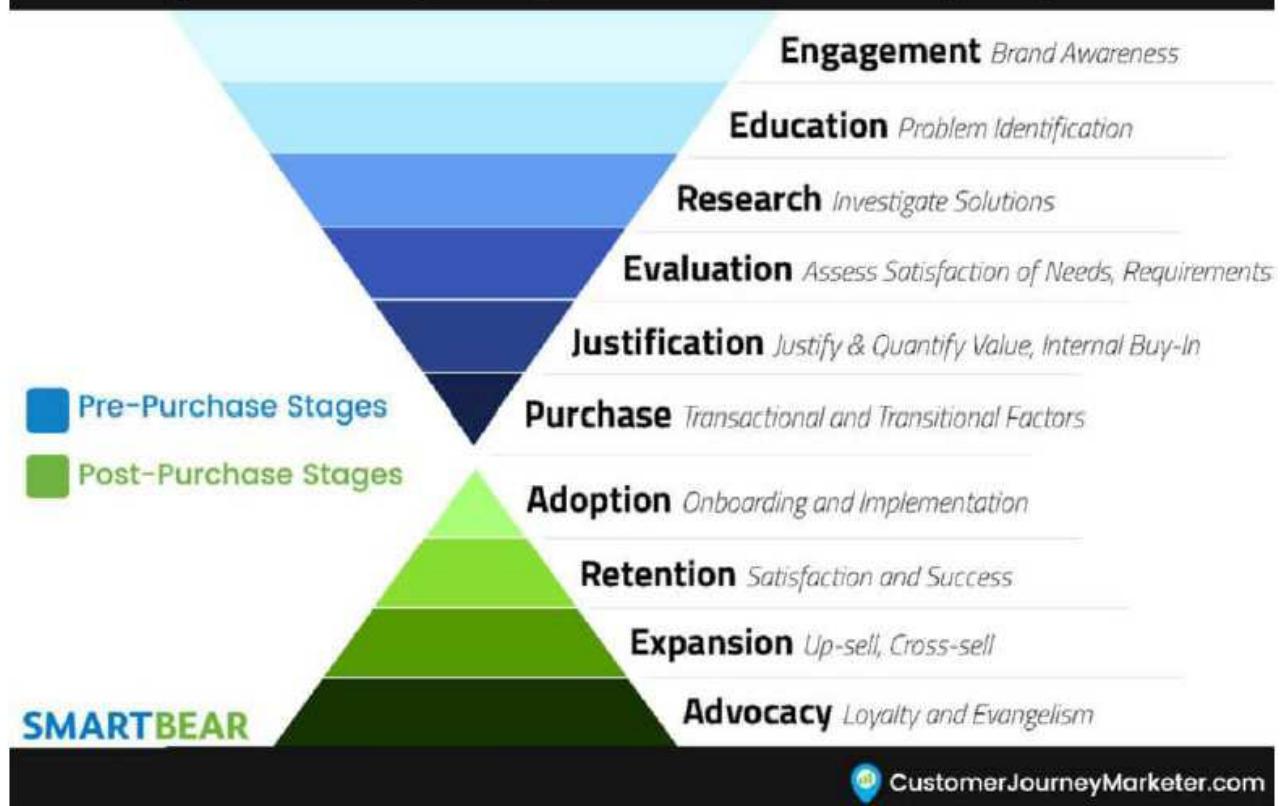


Figure 3: Digital Marketing Customer Stages Model

[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
```

```
bibtext _ label error
```

```
=====
```

```
bibtext space label error
```

```
=====
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
[2017-12-10 13.52.04] 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.
./README.yml
35:81     error      line too long (1061 > 80 characters)  (line-length)
35:1060   error      trailing spaces  (trailing-spaces)
65:22     error      no new line character at the end of file  (new-line-at-end-of-file)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Ashley Miller
hid: 329
paper1: 100% Oct 22
paper2: 100% Nov 6
project: 100% Dec 4
```

yamlcheck

---

wordcount

---

13

wc 329 project 13 8880 report.tex  
wc 329 project 13 9296 report.pdf  
wc 329 project 13 1750 report.bib

find "

---

32: The customer experience journey has been largely explored from a psychological and behavioral standpoint \cite{Stoicescu2015}.

Dating back to the near 1800s, marginal and expected utility of actions were detailed by Nicholas Bernoulli, among others, to better understand how purchasing decisions were made \cite{Stoicescu2015}. From related work that followed in this field of studying behavioral economics, research has shown that purchasing decisions are not linear and at times, are not even rational as cognitive, emotional, and social factors can all play a role into how a customer makes a purchasing decision \cite{Stoicescu2015}. As described by Stoicescu, the reason why researchers started to study purchasing behavior was due to the "diversification of need" \cite{Stoicescu2015}.

53: The way \textit{customer experience journey} is defined can differ by industry, product, and even by place. While past work has defined the customer experience journey as the process of purchasing a product or service, in today's landscape, it has become more than that. The Harvard Business Review would define the customer experience journey as the "sum-totality of how customers engage with your company or brand, not just in a snapshot of time, but throughout the entire arc of being a customer" \cite{Richardson2010}. Traditionally, the customer experience journey and buying process were used interchangeably where a customer moves through a decision making process. Some key areas that were highlighted in a typical customer experience journey include:

86: While advertising and marketing methods go as far back as the 1800s where customer lists were used to determine how individuals could be influenced via direct mailing efforts, digital marketing

has only come to be with the creation of the internet \cite{Couldry2014}. The internet has created opportunity for brands to directly connect with customers and likewise, for customers to engage with brands in a myriad of ways in the digital space \cite{Edelman2010}. While varying definitions of digital marketing exist, it is often categorized as a subset of traditional marketing where the “use of digital technologies create an integrated, targeted, and measurable communication” to not only attract potential customers but engage with current ones for retention and loyalty purposes \cite{Grishikashvili2014}. Digital marketing became even more prevalent in the 2000s as companies such as Google, Yahoo, and Facebook provided opportunity to deliver ads at an individual level based on demographic and behavioral characteristics \cite{Couldry2014}. Other data collection firms offered the ability to track users across the web space to see which pages were viewed, clicked, and time explored to help further understand the experience of a customer across the web space \cite{Couldry2014}. With these advances in technology and understanding, the customer experience journey began to also transform along with the changes in advertising from traditional to more digital.

- 91: One could argue that marketing data has been \textit{big} for quite some time given the sheer number of people exposed to efforts typically exceeds millions \cite{Couldry2014}. However, what has changed in this space is marketing’s increased use of digital technologies to reach potential customers in the digital landscape across various channels including search, display, social media, email, etc. \cite{Kannan2017}. For companies to be successful in utilizing big data to inform digital marketing efforts and to better track and enhance the customer experience journey, incorporating the necessary technologies and talent is a must along with shifting the organizational culture to making data driven decisions \cite{Grishikashvili2014}. This process can be difficult as an individual user alone can generate “billions of data signals” and attempting to understand which ones may be tied directly to a product or brand’s marketing efforts can be a daunting and overwhelming task \cite{Couldry2014}. However, there are a few well-known companies that have started the shift in tracking the customer experience journey through big data analytics and applications. We will examine key industries that have leveraged this knowledge and insight to better understand and enhance their relationship with customers.

- 107: eBay is a website that also started in 1995 which offered a unique opportunity to bring together buyers and sellers in the

online space \cite{Ebay2017}. Sellers could place their items online where buyers could potentially place bids, similar to if they participated in a live auction, where the item may go to the highest bidder. This allowed for others around the globe to connect and purchase items directly from another person while paying for the item through online methods. It is estimated today that there are nearly 180 million buyers and sellers and nearly 250 million search inquiries made per day on eBay \cite{Lui2012}. Like Amazon, eBay seeks to understand and tailor the online experience for customers through the use of big data in a multitude of ways as eBay itself states “understanding the customer is key” \cite{Sararn2014}. Various methods utilized to better understand and tailor the customer experience journey include:

- 110: \item \textbf{Web Page Metrics to Inform Layout}: It is estimated that among eBay customers, there is “100 million hours of interactions collected per month” \cite{Sararn2014}. Through an extensive number of experiments and A/B testing, eBay is able to optimize the web experience for customers. From their big data analytics, they can find preferred layouts of web pages which can customize anything from navigation feature to the size of photos displayed on the screen \cite{Akter2016}.
- 146: As others have stated about healthcare related data and reporting, “the problem in healthcare is not lack of data, but the unstructured nature of its data” \cite{Marr2016b}. Apixio, a cognitive computing firm based in California, wanted to take on the challenge of making unstructured healthcare related data available and easier to use in order to better aid decision making in patient treatment \cite{Marr2016b}. Their work involved taking clinical charts of patients and combining them with notes from physicians, test results, and even hospital stays to develop a more complete picture about an individual \cite{Marr2016b}. From there, Apixio was able to provide benefits based on this big data process:
- 165: Caesars has described their customer relationship optimization process as utilizing a “data-driven and closed loop approached to deliver a personalized experience” \cite{Welch2012}. A few ways they have implemented this include:
- 185: In just one day’s time, it is estimated that there are nearly 42,000 commercial flights and 2.5 million passengers \cite{Administration2017}. From purchasing a ticket, to taking a flight, and (hopefully) receiving their checked baggage at their

final destination, airlines collect a wealth of information on their customers throughout their flying experience \cite{Noyes2014}. When looking at key attributes that are analyzed and down to an individual level, airlines collect information about purchase history, arrival, departure cities, and dates, in-flight food choices, connecting cities, travel companions, as well as miles and credit card points earned and used \cite{Exastax2017}. While airlines have succeeded in collecting this data, using it to better enhance the customer experience journey is still a work in progress \cite{Noyes2014}. Those who work in the travel software environment and frequently provide products and services to those in the airline community to better understand their data even state they have "not seen a single major airline with an integrated big data business solution" \cite{Noyes2014}. With that in mind, highlights from major airline players are explored even though full development of utilizing big data may still be on-going in this industry.

- 202: These examples are a select few to showcase how companies can better understand and further the customer experience journey by leveraging big data. The average customer is presented with more choices today than ever before \cite{Meyer2007}. With this, companies today have to be more strategic to get the attention, time, and loyalty of customers to remain in the marketplace. Doing so can provide many advantages to both companies and customers as they utilize big data to better understand the customer experience. As Rawson et al state: "companies that excel in delivering journeys tend to win in the market" \cite{Rawson2013}. Trends presented showcase how big data can provide big benefit:
- 205: \item \textbf{Retention of Customers}: It is estimated that "acquiring a new customer can be between five and 25 times more expensive than retaining an existing one" \cite{Gallo2014}. Utilizing big data to predict when customers may close accounts can help to inform company efforts and ultimately prevent potential revenue loss if they can keep existing customers. American Express showed that by using big data and predictive analytics, the company could identify these customers sooner versus wait until the customer is already lost.
- 217: \item \textbf{Consumer Behavior Volatility} Not all decisions made by customers are rational ones and there can be a variety of factors in play that big data can not track \cite{Stoicescu2015}. As further detailed in other work "people do not behave like robots," so even when all the variables are optimized, outside

forces beyond the control of a company could influence choice along with a customer's own emotions which big data doesn't always include \cite{Richardson2010}.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

34: However, with diversification also comes complexity. The more choices a customer is given, the harder it can become for them to make a decision \cite{Stoicescu2015}. With every product choice, there also is an opportunity for interaction or \textit{touchpoints} along this customer experience journey \cite{Meyer2007}. These series of touchpoints can occur through a variety of ways and the time frame in which they take place can also vary greatly by the product or service being offered and to which audience. In figure \ref{f:Customer Experience Journey} example of a customer setting up utilities after the purchase of a new home and the multiple touchpoints they may encounter along their journey \cite{Rawson2013}.

36: \begin{figure}[ht!]

37: \centering\includegraphics[width=\columnwidth]{example.jpg}

38: \caption{Customer Experience Journey  
Touchpoints}\cite{Rawson2013}\label{f:Customer Experience Journey}

41: However, not all touchpoints are created equal \cite{Meyer2007}.

There are some touchpoints that every customer may have to go through to get to the next step in the process and others that will produce a more valuable action, such as a purchase

\cite{Meyer2007}. There are further questions today that did not exist in years past due to the advances in technology and how that affects customer behavior \cite{Kannan2017}. These advances in technology not only could influence customer behavior but also provide companies direction on which products they should produce, where these products should be placed, what price point is most optimal and how should they properly promote a particular product to their audience \cite{Kannan2017}. Big data and analytics can provide opportunity to inform the promotional piece as companies have utilized this feature to provide personalized and relevant content along the customer journey as defined in figure \ref{f:Customer Journey} \cite{Stoicescu2015}.

43: \begin{figure}[ht!]  
44: \centering\includegraphics[width=\columnwidth]{customerjourney.jpg}  
45: \caption{Customer Experience  
Journey}\cite{Stoicescu2015}\label{f:Customer Journey}

69: While the list of questions could be endless the intent is to move customers through this purchase decision process so companies create loyal customers and advocates \cite{Kannan2017}. However, that model is evolving with the shift to a multi-prong outreach approach via digital and non-digital methods \cite{DeAsi2017}. A longer customer experience journey is outlined in figure 1 as a customer can enter at any stage in the process. Pre and post purchase measures can be collected, stored, and analyzed at any point along the way as shown in figure \ref{f:Digital Marketing} \cite{DeAsi2017}.

71: \begin{figure}[ht!]  
72: \centering\includegraphics[width=\columnwidth]{digitalmarketing.jpg}  
73: \caption{Digital Marketing Customer Stages  
Model}\cite{DeAsi2017}\label{f:Digital Marketing}

figures 3  
tables 0  
\includegraphics 3  
labels 3  
refs 3  
floats 3

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

68: While the list of questions could be endless the intent is to move customers through this purchase decision process so companies create loyal customers and advocates \cite{Kannan2017}. However, that model is evolving with the shift to a multi-prong outreach approach via digital and non-digital methods \cite{DeAsi2017}. A longer customer experience journey is outlined in figure 1 as a customer can enter at any stage in the process. Pre and post purchase measures can be collected, stored, and analyzed at any point along the way as shown in figure \ref{f:Digital Marketing} \cite{DeAsi2017}.

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

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 Application Using MQTT and Raspberry Pi Robot Car

Arnav Arnav

Indiana University Bloomington  
Bloomington, Indiana 47408, USA  
aarnav@iu.edu

## ABSTRACT

As the number of connected edge devices increases there is a need for fast communication between these devices, and to analyse the data collected by these devices, which is made possible by the use of a scalable lightweight communication protocol such as MQTT, which is easy to use, data agnostic, and application independent. We look at one such application of the protocol, to control a robot car remotely, over wireless network, navigating with the help of a raspberry pi camera on the car.

## KEYWORDS

i523, HID201, Edge Computing, Raspberry Pi, MQTT, Robot Car, IoT

## 1 INTRODUCTION

As the number of edge devices increases, and sensor networks become more and more common in Internet of Things (IoT) applications, the need arises to allow these resource constrained devices to communicate with each other in a power efficient and secure manner. In many cases these devices may not be able to process traditional HTTP requests efficiently, and as the number of devices increases, sending an HTTP request to each of the devices in order to get data may not be efficient [3][11].

Message Queue Telemetry Transport (MQTT) is a lightweight machine to machine (M2M) messaging protocol, that uses a client/server based publish/subscribe model and is ideal for IoT applications. The protocol has been designed on top of TCP/IP protocol for us in situations where network bandwidth and available memory are limited [39][22]. The Eclipse Paho Project currently provides support for MQTT [5]. MQTT clients are available for various languages like Python, C, and Lua.

We look at one such application here that uses MQTT for communication between a raspberry pi and a desktop. The raspberry pi controls the stepper motors of the robot car according to the message it receives over mqtt, and drives the car accordingly. Another program running on the raspberry pi uses the raspberry pi onboard camera to capture pictures and send them back to the desktop to help in navigation. Thus we create a simple robot car that can be used remotely for monitoring purposes. The robot car can be controlled from anywhere in the world, as long as both the controlling device (desktop) and the raspberry pi can connect to the MQTT broker.

We can use multiple such cars and controlling devices to control the cars independently or from a common device to drive multiple cars together, thus controlling a swarm of cars. As these cars may be using different platforms like raspberry pi or arduino, Using MQTT allows us to write the controller program independent of the subscriber programs running on the different robot cars and

even in different languages. All that is needed to control a car is that the subscriber can understand the messages sent by the controller.

## 2 RELATED WORK

There have been many edge computing applications that involve robot cars or swarm of cars.

[28] provides an example of a raspberry pi car that uses distance sensor, and face detection on the raspberry pi 2. The car is controlled over wifi and is built using the GoPiGo robot car kit [14]

Zheng Wang used raspberry pi in [38] to build a sophisticated self driving car that can detect stop signs and traffic signals and drive appropriately on a small test track. The car has a camera and a distance sensor that stream data to a TCP server running on a desktop. The system uses Haar Cascades provided in opencv to detect objects like stop signs and traffic signals and a trained neural network which uses the image to predict the direction in which the car should move. The distance is calculated using the image from the raspberry pi camera with the help of a monocular vision method proposed by Chu, Ji, Guo, Li and Wang in 2004 [16].

As the part of Eclipse IoT open challenge [2] built a robot car that is controlled using the Constrained Application Protocol (CoAP) which snaps images and communicates the images over MQTT

OpenHAB provides a vendor neutral platform that allows users to integrate various home automation systems and provides an application interface to control those devices [25]. It allows integration of various devices with MQTT.

The FloodNet project at University of Southampton [12] aims at "providing a pervasive, continuous embedded monitoring presence". The system is intelligent and obtains "environmental self-awareness and resilience to ensure robust transmission of data", ensuring data quality and allowing exploration of environments in new ways. The project uses MQTT for communicating data from the sensors on field to visualization and simulation applications.

As a part of IBM's Extreme Blue projects, Say it Sign it [32] is a sophisticated, innovative speech to sign language translation system. The application uses speech recognition and renders an avatar that signs the corresponding words in British sign Language, using MQTT and microbroker for communication.

## 3 TECHNOLOGIES AND HARDWARE

The project uses MQTT to communicate between a controller running on a desktop and a raspberry pi that drives the robot car with the help of stepper motors. We describe these technologies in detail.

### 3.1 MQTT

MQTT works via a publish-subscribe model that contains 3 entities: (1) a publisher, that sends a message, (2) a broker, that maintains

queue of all messages based on topics and (3) multiple subscribers that subscribe to various topics they are interested in [29].

This allows for decoupling of functionality at various levels. The publisher and subscriber do not need to be close to each other and do not need to know each others identity. They need only to know the broker, as the publisher and the subscribers do not have to be running either at the same time nor on the same hardware [19].

MQTT implements a hierarchy of topics that are related to all messages. These topics are recognised by strings separated by a forward-slash (/), where each part represents a different topic level. This is a common model introduced in file systems but also in internet URLs.

A topic looks therefore as follows: *topic-level0/topic-level1/topic-level2*.

All subscribers subscribe to different topics via the broker. Subscribing to *topic-level0* allows the subscriber to receive all messages that are associated with topics that start with *topic-level0*.

This is different from traditional message queues as the message is forwarded to multiple subscribers, and allows for a more flexible approach with the help of topics [19]. The basic steps in an MQTT client application include connecting to the broker, subscribing to some topics, waiting for messages and performing the appropriate action when a certain message is received [39].

MQTT allows the publisher and subscriber to respond to messages with the help of callbacks that are executed on different events, in a non-blocking manner. The paho-mqtt package for python provides callbacks methods like `on-connect()`, `on-message()` and `on-disconnect()`, which are fired when the connection to the broker is complete, a message is received from the broker, and when the client is disconnected from the broker respectively. These methods are used in conjunction with the `loop-start()` and `loop-end()` methods which start and end an asynchronous loop that listens for these events and fires the relevant callbacks, allowing the clients to perform other tasks [6].

MQTT has been designed to be flexible and options are provided to easily change the quality of service (QoS) as required by the application. Three basic levels of QoS are supported by the protocol, Atmost-once (QoS level 0), Atleast-once (QoS level 1) and Atmost-once (QoS level 2) [20][6].

The QoS level of 0 can be used in applications where some dropped messages may not affect the application. Under this QoS level, the broker forwards a message to the subscribers only once and does not wait for any acknowledgement [20] [6].

The QoS level of 1 can be used in situations where the delivery of all messages is important and the subscriber can handle duplicate messages. Here the broker keeps on resending the message to a subscriber after a certain timeout until the first acknowledgement is received. A QoS level of 2 should be used in cases where all messages must be delivered and no duplicate messages should be allowed. In this case the broker sets up a handshake with the subscriber to check for its availability before sending the message [20] [6].

The MQTT specification uses TCP/IP to deliver the messaged to the subscribers, but it does not provide any form of security by default to make it useful for resource constrained IoT devices. “It allows the use of username and password for authentication,

but by default this information is sent as plain text over the network, making it susceptible to man-in-the middle attacks” [27] [21]. Therefore, in sensitive applications some form of additional security measures are recommended which may include network layer security with the use of Virtual Private Networks (VPNs), Transport Layer Security, or application layer security [21].

Transport Layer Security (TLS) and Secure Sockets Layer (SSL) are cryptographic protocols that establish the identity of the server and client with the help of a handshake mechanism which uses trust certificates to establish identities before encrypted communication can take place [4]. If the handshake is not completed for some reason, the connection is not established and no messages are exchanged [21]. “Most MQTT brokers provide an option to use TLS instead of plain TCP and port 8883 has been standardized for secured MQTT connections” [27].

Using TLS/SSL security however comes at an additional cost. If the connections are short-lived then most of the time can be spent in the handshake itself, which may take up few kilobytes of bandwidth. In case the connections are short-lived, temporary session IDs and session tickets can be used to resume a session instead of repeating the handshake process. If the connections are long term, the overhead of the handshake is negligible and TLS/SSL security should be used [27][21].

Although MQTT protocol itself does not include authorization, many MQTT brokers include authorization as an additional feature [4]. OAuth2.0 uses JSON Web Tokens which contain information about the token and the user and are signed by a trusted authorization server [10].

When connecting to the broker this token can be used to check whether the client is authorised to connect at this time or not. Additionally the same validations can be used when publishing or subscribing to the broker. The broker may use a third party resource such as LDAP (lightweight directory access protocol) to look up authorizations for the client [10]. Since there can be a large number of clients and it can become impractical to authorize everyone, clients may be grouped and the authorizations may be checked for each group [4].

MQTT allows easy integration with other services, that have been designed to process this data.

Apache storm is a distributed processing system that allows real time processing of continuous data streams, much like Hadoop works for batch processing [1]. Apache storm can be easily integrated with MQTT as shown in [36] to get real time data streams and allow analytics and online machine learning in a fault tolerant manner [42].

ELK stack (elastic-search, logstash and kibana) is an open source project designed for scalability which contains three main software packages, the *elastic-search* search and analytics engine, *logstash* which is a data collection pipeline and *kibana* which is a visualization dashboard [7]. Data from an IoT network can be collected, analysed and visualized easily with the help of the ELK stack as shown in [34] and [33].

MQTT broker services can be utilized for enterprise and production environments. EMQ (Erlang MQTT Broker) provides a highly scalable, distributed and reliable MQTT broker that can be used in enterprise-grade applications [9].

## 3.2 Raspberry Pi

The raspberry pi is a credit card sized development board that was developed by Eben Upton with the goal to create a low cost device that can be used for education and prototyping [26]. Since its creation the board has been adapted for various different projects by educators hobbyists and in the industry [31]. The board is developed as open hardware except for the Broadcom chip that controls the main components of the board, and most raspberry pi projects are available openly with detailed documentation.

The board's Broadcom system on chip consists of an ARM processor and it can be used just like a normal computer by connecting a monitor, a keyboard and a mouse. The raspberry pi can communicate to other devices with the help of wifi and bluetooth and is capable of accessing the internet. All this put together makes the raspberry pi a very useful device [31].

The raspberry pi comes in various models, Model A+, which is one of the smallest form factors, raspberry pi2 Model B, raspberry pi3 Model B and Model B+ that have more gpio pins. The raspberry pi 3 Model B is the newest design and consists of on board wifi and bluetooth, eliminating the need to use usb wifi and bluetooth attachments. It has a 1.2 GHz ARM 8 microprocessor, 1 GB RAM, a dual core Videocore IV GPU, and 40 general purpose input and output (GPIO) pins. The board has an ethernet port and four USB ports and an HDMI port to connect to a monitor [18][17].

The raspberry pi Zero is the development board that has the smallest form factor. Even though the raspberry pi zero includes no ethernet or USB ports, and does not come with GPIO pins soldered on, its small size and cost effectiveness make it extremely useful in applications such as IoT where space is constrained [30].

The raspberry pi uses a micro SD card to boot and various operating systems, that support the ARM architecture can be used. The most common operating systems are Raspbian, a derivative of the Debian linux, and Pidora, a derivative of Fedora. There are other operating systems centered around using the raspberry pi for various purposes, like openELEC and RaspBMC, which make it easy to use raspberry pi as a multimedia center. For, users who want non-linux operating system, RISC OS may be a good choice. The raspberry pi foundation provides new users the opportunity to try out various operating systems with the help of their New Out Of The Box Software (NOOBS), which allows the users to pick which operating system they want to use [26].

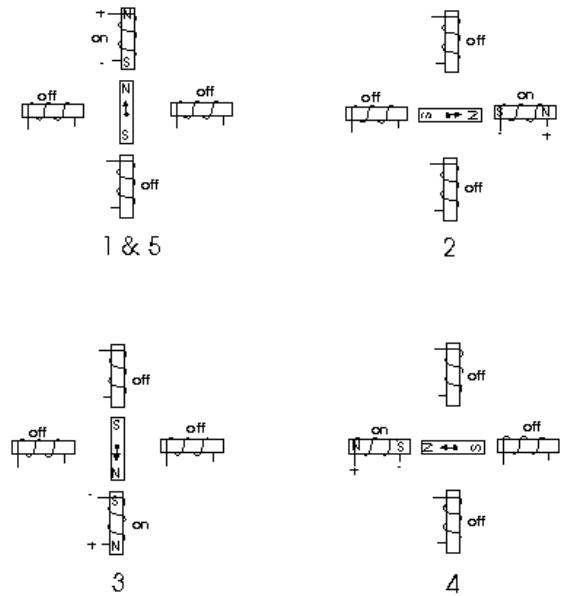
Various different shields are available for the raspberry pi that make it simple to connect to various peripherals, and extend the functionality of the raspberry pi, such as the GrovePi shield, provided by Dexter Industries, which allows simple interface with many digital and analog sensors and actuators provided by Dexter.

## 3.3 Stepper Motors

Stepper motors are brushless motors that divide the complete rotation into a number of parts known as steps. The motor consists of electromagnetic coils and a rotating core that aligns itself according to the combined magnetic effect of the coils. The stepper motor can move from one step to another and remain in a single step based on which coils are turned on. The torque of the motor can be increased or decreased with the current supplied to the coils, and the speed

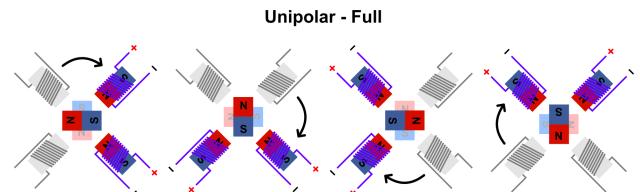
of rotation can be controlled by setting the time interval between switching the coils on and off [41].

Stepper motors can be controlled in various ways, depending on the application. Figure 1 shows how a stepper motor with a resolution of 90 degrees can be made to complete one full rotation. In practice however, the resolution (the degrees moved at each step) of most stepper motors is much higher. The process mentioned in figure 1 is known as half stepping [15].



**Figure 1: Working of a stepper motor : Full stepping using one coil at a time [15]**

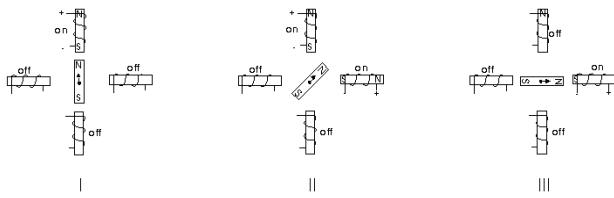
In the above method, only one coil is turned on at a time. This can be improved upon to get a higher torque. To get a higher torque, two adjacent coils are turned on at the same time, as shown in figure 2. This results in double the torque generated when using only one coil at a time [37].



**Figure 2: Working of a stepper motor : Full stepping using two coils at a time [37]**

With full stepping however, the transition between two consecutive steps is not very smooth. Therefore, a technique called Half

stepping is used, where two adjacent coils are turned on similar to full stepping, but between two steps one of the coils is turned off, so that the transition between steps is smooth. This results in a torque 70 percent of that generated in using full stepping with two coils turned on at the same time. This process is shown in figure 3 [15].



**Figure 3: Working of a stepper motor : Half stepping [15]**

For this project, the stepper motor 28BYJ-48, provided by Elegoo Industries is used. The motor is driven with the help of a ULN2003 motor driver. The motor is a unipolar stepper motor, with a five wire connection to the motor controller and can work with 5 and 12 Volts of DC power supply. When using Half stepping, the step angle of the motor is about 5.625 degrees per step, and when using full stepping the step angle is 11.25 degrees per step. The motor weighs 30 grams, and a gear ratio of 64:1 [13][35].

### 3.4 OpenCV

The Open Source Computer Vision library (openCV) is a library of functions aimed at real time computer vision and machine learning and providing a common infrastructure to allow fast progress in the field of computer vision and machine perception [40][23].

The library was originally built by Intel and is now maintained by Itseez and is available freely under open-source BSD License. The library was originally written for C++ but has been developed as cross platform library and supports Python , C++, MATLAB and Java [23]. for Python the library has been built on top of Numpy, a library that optimizes matrix and vector operations, and takes advantage of MMX and SSE instructions whenever possible. For C++ the library uses the Standard Template Library (STL) as its backbone.

The library has more than 2500 algorithms which include a combination of simple and advanced operations allowing a wide range of operations from edge detection, color detection to object detection, face detection and automatic video stabilization, and motion detection. The opencv-contrib which is an extension to the library built collaboratively by the community contains advanced algorithms that allow processing video in real time [23].

OpenCV is widely used in the industry by startups as well as well established organizations like Google, Yahoo, Microsoft, IBM and Intel [23].

OpenCV can be used to detect faces in real time. The Haar cascades function in the library allows detecting any kind of objects. The algorithm uses a series of simple classifiers to predict whether a given image has the desired object or not. After training on a large set of positive examples (images containing faces) and negative examples (images not containing faces), the algorithm learns

various classifiers, that classify different sections of the image in a manner similar to Adaboost algorithm. Only the portions of the image that are promising are analysed further by more detailed classifiers. This allows the algorithm to run in real time, and detect multiple objects [24][43]. Once the classifiers are learned, they can be stored in an XML file which can be used to classify new images. This allows users to obtain XML files available openly for classifiers trained to detect the required object and use them in their programs. OpenCV provides XML files for classifiers trained to detect faces and eyes.

The performance of the Haar cascades suffers however, when detecting objects in new images that are present in a different orientation than the ones used to train the classifiers. The classifiers may also fail to differentiate between the object that needs to be detected and similar objects if enough negative examples are not shown while training that include similar objects.

## 4 ARCHITECTURE

The solution includes two entities the raspberry pi and the desktop, each running two programs. The raspberry pi is connected to the robot car and the raspberry pi can drive the robot car according to the message it receives from the desktop.

There are two programs running on pi, controller stepper sub.py and video pub.py, and two programs running on the desktop, controller pub.py and video sub.py.

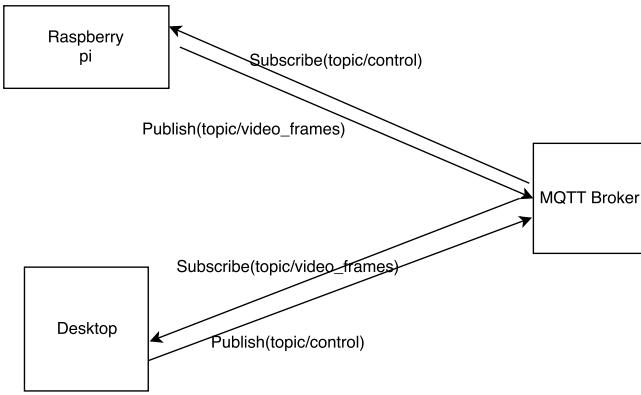
The programs on both the raspberry pi and the desktop connect to a common broker. The broker may be running on the desktop, or any other place, as long as the IP address of the broker is known. The IP address can be passed as a command line argument when running these programs.

The controller pub.py program running on the desktop continuously reads characters from the user and publishes them to the broker under the topic *topic/control*. The subscriber controller stepper sub.py running on the raspberry pi waits for these messages from the broker and when a message is received it uses the *on\_message()* callback to make the robot car move forward, move backward, turn left or turn right, using the half stepping technique described in the previous section.

For monitoring purposes, another program, video pub runs on the raspberry pi. This program uses the raspberry pi on board camera with the help of the picamera module and captures images. The images are converted to greyscale, and opencv is used to perform face detection using Haar Cascades. If a face is found, a box is drawn around the face in the image. The image is published to the broker under the topic *topic/video\_frames*. The video sub.py program running on the desktop subscribes to this topic on the broker and displays the images received. These images can be used for the navigation of the robot car remotely figure 4.

Using separate programs allow changing the functionality or replacing different parts of the program easily, while keeping the interface same. The program, controller sub.py, can be used if continuous rotation servo motors are used instead of the stepper motors without changing any other part of the application.

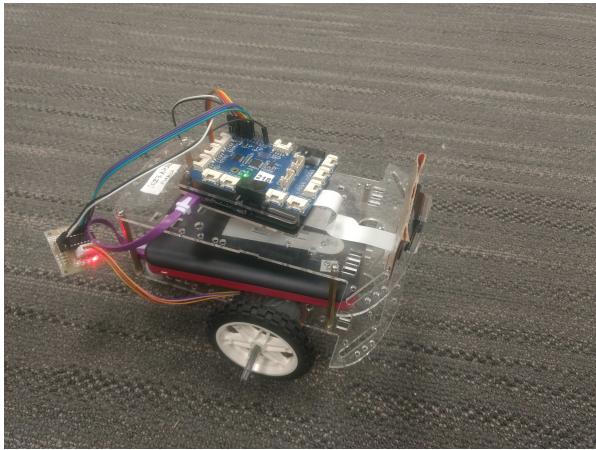
The programs can be run easily with the help of a Makefile as described in the next section



**Figure 4: Architecture of the Application**

## 5 RESULTS

This section covers the setup instructions for the project and the observations. The robot car that was built is shown in figure 5.



**Figure 5: Raspberry Pi Robot Car**

### 5.1 Setup Instructions

To run the application successfully on both the raspberry pi and the desktop, it must be ensured that all the required libraries are installed. A Makefile has been provided that can do this on both the raspberry pi and the desktop.

- First, the motors should be connected to the raspberry pi correctly. The program uses the raspberry pi GPIO pins, and assumes that for the left motor, the pins IN1, IN2, IN3, IN4 are connected to GPIO pins 7, 11, 13, and 15, and for the right motor, they are connected to GPIO pins, 8, 10, 12, 16, as shown in the connection diagram in figure 6
- On the raspberry pi, dependencies for openCV need to be installed. Since the openCV is not available in pip for the arm processor in raspberry pi, we it must be installed from

source. This takes a few hours on the raspberry pi. To complete the setup including installation of a MQTT client and opencv on the raspberry pi, clone the repository from github on the raspberry pi and navigate to the code folder, open the terminal and run the command

`make setup_pi`

- Next, install opencv and an MQTT client and MQTT broker on the desktop. For this, clone the repository from github, navigate into the code folder and run the command  
`make setup_server`
- Note the IP address of the desktop so that we can connect to the MQTT server running on it. Connect the raspberry pi and the desktop on the same wireless network.
- To run the code on the desktop, run the command  
`make run_server IP=[IP address of the MQTT broker]`
- Finally to run the code on the raspberry pi, run  
`make run_pi IP=[IP address of the MQTT broker]`
- Now the raspberry pi car can be controlled by typing in W, A, S, or D keys on the desktop in the terminal where the program is running.
- The program can be stopped on both the raspberry pi and the desktop by running  
`make kill`

### 5.2 Observations

It was observed that the communication between the raspberry pi and the desktop controller application is pretty seamless. The robot car responds without any observable delays when the network is strong. When the network is weak, however, some delays may be observed. The delay becomes more evident in the case of the images sent by the raspberry pi back to the desktop when the network is not strong.

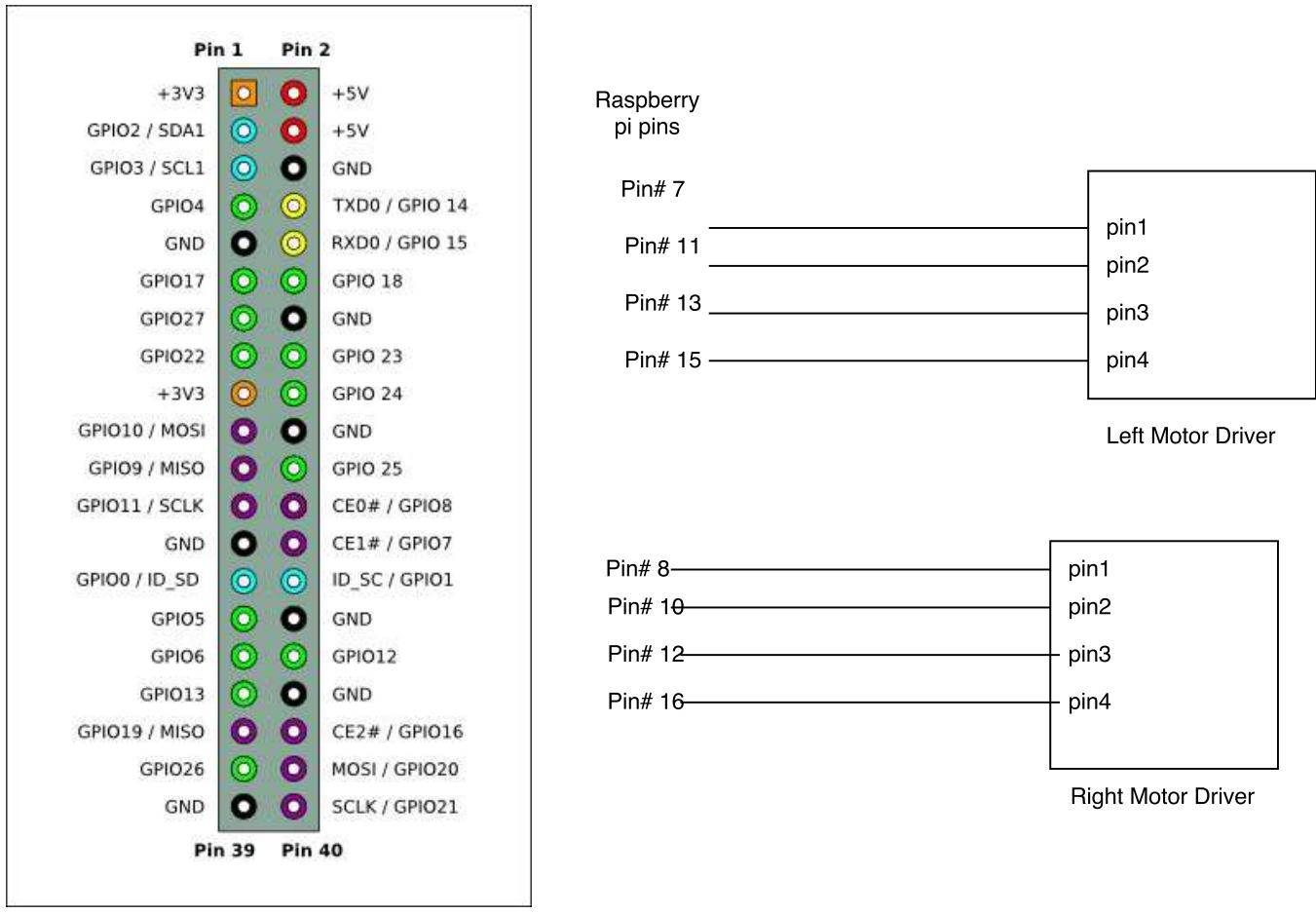
Using the stepper motors, it is difficult to set a how much a motor should turn when it receives a message. If the motor is not allowed to turn long enough, then between two messages the motor will be idle and if it is turned longer than the interval between two messages, there can be conflicts if in response to each of the messages the subscriber running on the raspberry pi tries to set a different step on the motor. Therefore, the movement can seem a little jerky at times.

However, this is not a problem with 360 degrees continuous servo motors. Since the continuous servo motors use pulse width modulation, the speed and direction of rotation can be controlled by sending a square wave with different duty cycles depending on the motor. Since, the motor can be stopped and started easily, there are no conflicts even if the motor is allowed to turn longer than the interval between two messages. However, the motor would respond to the two messages one after the other.

Thus the raspberry pi robot car can be successfully controlled over wifi using MQTT for communication

### 5.3 Improvements

The project can be improved in various ways. Firstly, even though the deployment with makefile is easy, installing opencv on raspberry pi takes around 4 hours. This can be avoided if we use docker for deployment on the raspberry pi. Two separate images would me



Raspberry pi 3 GPIO header

Figure 6: Connection Diagram [8]

needed however one for the processor on the desktop and another one for the arm 8 processor on the raspberry pi.

Machine learning can be incorporated, by collecting the images and the corresponding messages that were sent to the raspberry pi and use it to train a neural network, which could then be used to drive the robot car autonomously. This would be complicated however since car needs to be driven for a long time to get enough data for the neural network to perform well regardless of the surroundings.

Using Haar cascades for face detection leads to a problem that faces can be recognised only if they are resent in the image in the same orientation as that in the training examples. Therefore, it is challenging to recognise all faces in all orientations since it is not possible to train the classifier on images of different faces from all possible angles and rotations. A better option would be to use Convolutional Neural Networks, that help in improving accuracy for the purpose of object detection. Since training and running neural networks may be computationally expensive, it would be a good idea to run it on a server and not on the raspberry pi.

Many different sensors could be added to help improve the monitoring capability of the car, and get more information about the environment. If many controlling devices and cars are present, the cars may be controlled in groups and other functionality added to behave as a swarm of cars to complete tasks collaboratively.

## 6 CONCLUSION

MQTT is a fast and reliable data agnostic and platform independent protocol that allows communication between devices. Raspberry pi is small but powerful development board that allows users to build prototypes easily and can be used in various applications because of the significantly powerful arm 8 microprocessor. OpenCV is an open source library for computer vision that is optimised to perform operations on images efficiently and is commonly used in computer vision applications. All these technologies were used to build a robot car, controlled via MQTT over a wireless network. MQTT allows us to easily scale up the number of such cars if needed.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for giving the opportunity to work on this project and for providing the necessary hardware to complete the project.

The author would also like to thank the associate instructors of the class for their help and for answering questions on piazza which helped everyone.

## REFERENCES

- [1] apache. [n. d.]. apache storm. apache storm website. ([n. d.]). <http://storm.apache.org/>
- [2] bitreactive. 2015. The Raspberry Pi Eclipse IoT Car. bitreactive website. (March 2015). <http://www.bitreactive.com/remote-controlled-raspberry-pi-car-part-3-2/>
- [3] Paul Caponetti. 2017. Why MQTT is the Protocol of Choice for the IoT. xively.com blog website. (august 2017). <http://blog.xively.com/why-mqtt-is-the-protocol-of-choice-for-the-iot/>
- [4] Ian Craggs. 2013. MQTT security: Who are you? Can you prove it? What can you do? IBM developer works website. (march 2013). [https://www.ibm.com/developerworks/community/blogs/c565c720-fe84-4f63-873f-607d87787327/entry/mqtt\\_security?lang=en](https://www.ibm.com/developerworks/community/blogs/c565c720-fe84-4f63-873f-607d87787327/entry/mqtt_security?lang=en)
- [5] eclipse. [n. d.]. mqtt broker. eclipse mosquitto website. ([n. d.]). <https://mosquitto.org/>
- [6] eclipse paho. [n. d.]. Python Client - documentation. eclipse paho website. ([n. d.]). <https://www.eclipse.org/paho/clients/python/docs/>
- [7] elastic.io. [n. d.]. ELK stack. elastic.io website. ([n. d.]). <https://www.elastic.co/products>
- [8] eLinux.org. 2015. File:Pi-GPIO-header.png. elinux.org website. (July 2015). <https://elinux.org/images/5/5c/Pi-GPIO-header.png>
- [9] erlang mqtt. [n. d.]. erlang mqtt broker. wmqtt website. ([n. d.]). <http://emqtt.io/docs/v2/index.html>
- [10] hive mq. [n. d.]. MQTT Security Fundamentals: OAuth 2.0 & MQTT. hivemq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-security-fundamentals-oauth-2-0-mqtt>
- [11] hivemq. [n. d.]. intrewebsite mqtt. hivemq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-essentials-part-1-introducing-mqtt>
- [12] University of Southampton IAM group. 2005. FloodNet. IAM group website. (April 2005). <http://www.iam.ecs.soton.ac.uk/projects/297.html>
- [13] Elegoo Industries. 2017. Elegoo 5 sets 28BYJ-48 5V Stepper Motor and ULN2003 Motor Driver Board for Arduino. elegoo industries website. (2017). <https://www.elegoo.com/product/elegoo-5-sets-28byj-48-5v-stepper-motor-uln2003-motor-driver-board-for-arduino/>
- [14] Dexter Industries. 2017. GoPiGo Build and Program Your Own Robot. dexter industries website. (2017). <https://www.dexterindustries.com/gopigo3/>
- [15] Images Scientific Instrumentation. 2017. How Stepper Motors Work. imagesco.com website. (2017). <http://www.imagesco.com/articles/picstepper/02.html>
- [16] Chu Jiangwei, Ji Lisheng, Guo Lie, Wang Rongben, et al. 2004. Study on method of detecting preceding vehicle based on monocular camera. In *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 750–755.
- [17] jwatson. 2016. Raspberry Pi Models Comparison Chart Poster. element14 community website. (June 2016). <https://www.element14.com/community/docs/DOC-82195/l/raspberry-pi-models-comparison-chart-poster-free-download>
- [18] makershed.com. 2016. Raspberry pi comparison chart. makershed.com website. (2016). <https://www.makershed.com/pages/raspberry-pi-comparison-chart>
- [19] Hive mq. [n. d.]. MQTT Essentials Part 2: Publish & Subscribe. HiveMQ website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-essentials-part2-publish-subscribe>
- [20] Hive MQ. [n. d.]. MQTT Essentials Part 6: Quality of Service 0, 1 & 2. Hivemq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-essentials-part-6-mqtt-quality-of-service-levels>
- [21] Hive MQ. [n. d.]. MQTT Security Fundamentals: TLS / SSL. hivemq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-security-fundamentals-tls-ssl>
- [22] Mqtt. [n. d.]. Mqtt official website. mqtt official website. ([n. d.]). <http://mqtt.org/>
- [23] OpenCV. 2017. About. opencv.org website. (2017). <https://opencv.org/about.html>
- [24] OpenCV. 2017. Face Detection using Haar Cascades. opencv website. (August 2017). [https://docs.opencv.org/3.3.0/d7/d8b/tutorial\\_py\\_face\\_detection.html](https://docs.opencv.org/3.3.0/d7/d8b/tutorial_py_face_detection.html)
- [25] OpenHAB. 2017. What is openHAB? openhab website. (November 2017). <https://www.openhab.org/introduction.html>
- [26] opensource.com. 2015. What is a Raspberry Pi. opensource.com website. (March 2015). <https://opensource.com/resources/raspberry-pi>
- [27] Todd Ouska. 2016. Transport-level security tradeoffs using MQTT. iot design website. (February 2016). <http://iotdesign.embedded-computing.com/guest-blogs/transport-level-security-tradeoffs-using-mqtt/>
- [28] pythonprogramming.net. 2014. Robotics with Python Raspberry Pi and GoPiGo Introduction. pythonprogramming.net. (April 2014). <https://pythonprogramming.net/robotics-raspberry-pi-tutorial-gopigo-introduction/>
- [29] random nerds tutorial. [n. d.]. What is MQTT and How It Works. random nerds website. ([n. d.]). <https://randomnerdtutorials.com/what-is-mqtt-and-how-it-works/>
- [30] raspberrypi.org. 2015. Raspberry Pi Zero: the 5 dollar computer. raspberrypi.org. (November 2015). <https://www.raspberrypi.org/blog/raspberry-pi-zero/>
- [31] raspberrypi.org. 2015. What is a Raspberry pi. raspberrypi.org website. (May 2015). <https://www.raspberrypi.org/help/what-%20is-a-raspberry-pi/>
- [32] IBM research. 2007. IBM Research Demonstrates Innovative 'Speech to Sign Language' Translation System. IBM website. (September 2007). <http://www-03.ibm.com/press/us/en/pressrelease/22316.wss>
- [33] smart factory. 2016. MQTT and Kibana fit! Open source Graphs and Analysis for IoT. smart factory website. (May 2016). <https://smart-factory.net/mqtt-and-kibana-open-source-graphs-and-analysis-for-iot/>
- [34] smart factory. 2016. Storing IoT data using open source. MQTT and ElasticSearch fit! Tutorial. smart factory website. (october 2016). <https://smart-factory.net/mqtt-elasticsearch-setup/>
- [35] Stan. 2014. 28BYJ-48 Stepper Motor with ULN2003 driver and Arduino Uno. 42 bolts website. (March 2014). <http://42bots.com/tutorials/28byj-48-stepper-motor-with-uln2003-driver-and-arduino-uno/>
- [36] Apache storm. [n. d.]. Storm MQTT Integration. Apache storm website. ([n. d.]). <http://storm.apache.org/releases/1.1.0/storm-mqtt.html>
- [37] Built to spec. 2015. Understanding Stepper Motors Part I fit!! A Basic Model. built-to-spec.com website. (October 2015). <http://www.built-to-spec.com/blog/2012/04/09/understanding-stepper-motors-part-i-a-basic-model/>
- [38] Zheng Wang. 2015. Self Driving RC Car. Zheng Wang wordpress website. (August 2015). <https://zhengludwig.wordpress.com/projects/self-driving-rc-car/>
- [39] Wikipedia. 2017. MQTT – Wikipedia, The Free Encyclopedia. (November 2017). <https://en.wikipedia.org/w/index.php?title=MQTT&oldid=808683219> [Online; accessed 6-November-2017].
- [40] Wikipedia. 2017. OpenCV – Wikipedia, The Free Encyclopedia. (2017). <https://en.wikipedia.org/w/index.php?title=OpenCV&oldid=811519079> [Online; accessed 4-December-2017].
- [41] Wikipedia. 2017. Stepper motor – Wikipedia, The Free Encyclopedia. (2017). [https://en.wikipedia.org/w/index.php?title=Stepper\\_motor&oldid=811220740](https://en.wikipedia.org/w/index.php?title=Stepper_motor&oldid=811220740) [Online; accessed 4-December-2017].
- [42] Wikipedia. 2017. Storm (event processor) – Wikipedia, The Free Encyclopedia. (2017). [https://en.wikipedia.org/w/index.php?title=Storm\\_\(event\\_processor\)&oldid=80871136](https://en.wikipedia.org/w/index.php?title=Storm_(event_processor)&oldid=80871136) [Online; accessed 6-November-2017].
- [43] Wikipedia. 2017. ViolaFit!Jones object detection framework – Wikipedia, The Free Encyclopedia. (2017). <https://en.wikipedia.org/w/index.php?title=Viola%20%80%93Jones.object.detection.framework&oldid=808683512> [Online; accessed 4-December-2017].

```
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 apache-storm
Warning--empty year in eclipse-mosquitto
Warning--empty year in python-paho-mqtt
Warning--empty year in elk-stack
Warning--empty year in erlang-mqtt-broker
Warning--empty year in hivemq-security-oauth
Warning--empty year in hivemq-website
Warning--empty publisher in monocular
Warning--empty address in monocular
Warning--empty year in hivemq-details
Warning--empty year in hivemq-qos
Warning--empty year in mqtt-sec-ssl
Warning--empty year in mqtt-official
Warning--empty year in how-mqtt-works
Warning--empty year in apache-storm-mqtt
(There were 15 warnings)
```

```
bibtext _ label error
```

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

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

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

```
=====
[2017-12-10 13.47.11] 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: ""
```

```
Typesetting of "report.tex" completed in 1.0s.
```

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

```
name: Arnav, Arnav
hid: 201
paper1: 20th Oct 2017 100%
paper2: 100%
project: 100%
```

```
yamlcheck
```

```
wordcount
```

```
7
wc 201 project 7 4877 report.tex
wc 201 project 7 5274 report.pdf
wc 201 project 7 1807 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

---

- 228: Figure \ref{f:stepper1} shows how a stepper motor with a resolution of 90 degrees can be made to complete one full rotation. In practice however, the resolution (the degrees moved at each step) of most stepper motors is much higher. The process mentioned in figure \ref{f:stepper1} is known as half stepping \cite{stepper1}.
- 230: \begin{figure}[!ht]
- 231: \centering\includegraphics[width=\columnwidth]{images/stepper1.pdf}
- 232: \caption{Working of a steper motor : Full stepping using one coil at a time \cite{stepper1}}\label{f:stepper1}
- 235: In the above method, only one coil is turned on at a time. This can be improved upon to get a higher torque. To get a higher torque, two adjacent coils are turned on at the same time, as shown in figure \ref{f:stepper2}. This results in double the torque generated when using only one coil at a time \cite{stepper2}.
- 237: \begin{figure}[!ht]
- 238: \centering\includegraphics[width=\columnwidth]{images/stepper2.pdf}
- 239: \caption{Working of a steper motor : Full stepping using two coils at a time \cite{stepper2}}\label{f:stepper2}
- 242: With full stepping however, the transition between two consecutive steps is not very smooth. Therefore, a technique called Half stepping is used, where two adjacent coils are turned on similar to full stepping, but between two steps one of the coils is turned off, so that the transition between steps is smooth. This results in a torque 70 percent of that generated in using full stepping with two coils turned on at the same time. This process is shown in figure \ref{f:stepper3} \cite{stepper1}.
- 244: \begin{figure}[!ht]
- 245: \centering\includegraphics[width=\columnwidth]{images/stepper3.pdf}
- 246: \caption{Working of a steper motor : Half stepping \cite{stepper1}}\label{f:stepper3}
- 276: For monitoring purposes, another program, video\\_pub runs on the raspberry pi. This program uses the raspberry pi on board camera with the help of the picamera module and captures images. The images are converted to greyscale, and opencv is used to perform face detection using Haar Cascades. If a face is found, a box is drawn around the face in the image. The image is published to the broker under the topic {\em topic/video\\_frames}. The

video\\_sub.py program running on the desktop subscribes to this topic on the broker and displays the images received. These images can be used for the navigation of the robot car remotely figure \ref{f:arch}.

```

278: \begin{figure}[!ht]
279: \centering\includegraphics[width=\columnwidth]{images/architectur
e.pdf}
280: \caption{Architecture of the Application}\label{f:arch}
287: This section covers the setup instructions for the project and
the observations. The robot car that was built is shown in figure
\ref{f:car1}. %and figure \ref{f:car2}
289: \begin{figure}[!ht]
290: \centering\includegraphics[width=\columnwidth]{images/car1.pdf}
291: \caption{Raspberry Pi Robot Car}\label{f:car1}
295: \%begin{figure}[!ht]
296: %
\centering\includegraphics[width=\columnwidth]{images/car2.pdf}
297: \% \caption{Raspberry Pi Robot Car}\label{f:car2}
308: \ref{f:connection}
311: \centering\includegraphics[width=\textwidth]{images/connection.pd
f}
312: \caption{Connection Diagram \cite{rpi-
pinout}}\label{f:connection}
```

figures 6  
tables 0  
\includegraphics 7  
labels 7  
refs 6  
floats 6

```

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

---

```
311: \centering\includegraphics[width=\textwidth]{images/connection.pd  
f}
```

passed: False

below\_check

---

WARNING: figure and above may be used improperly

235: In the above method, only one coil is turned on at a time. This can be improved upon to get a higher torque. To get a higher torque, two adjacent coils are turned on at the same time, as shown in figure \ref{f:stepper2}. This results in double the torque generated when using only one coil at a time \cite{stepper2}.

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 apache-storm  
Warning--empty year in eclipse-mosquitto  
Warning--empty year in python-paho-mqtt  
Warning--empty year in elk-stack  
Warning--empty year in erlang-mqtt-broker  
Warning--empty year in hivemq-security-oauth  
Warning--empty year in hivemq-website  
Warning--empty publisher in monocular  
Warning--empty address in monocular  
Warning--empty year in hivemq-details  
Warning--empty year in hivemq-qos  
Warning--empty year in mqtt-sec-ssl  
Warning--empty year in mqtt-official  
Warning--empty year in how-mqtt-works
```

```
Warning--empty year in apache-storm-mqtt
(There were 15 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
```

```
[2017-12-10 13.51.09] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.1   L32    : [WundergroundOverview2017] undefined
p.1   L32    : 'TotPages' undefined
p.2   L42    : [NEST2017] undefined
p.2   L42    : [NESTRReddit2017] undefined
Command \starttoc has changed. Check if current package is valid.
File 'RASP3_1.jpg' not found.
./report.tex:67: Package pdftex.def Error: File 'RASP3_1.jpg' not found.
```

```
L67      :           \includegraphics[scale=.10]{RASP3_1.jpg}
LaTeXError: Package pdfTeX.def Error: File ‘RASP3_1.jpg’ not found.
```

# Face Detection and Recognition Using Raspberry Pi Robot Car

Mani Kumar Kagita

Indiana University

107 S. Indiana Avenue

Bloomington, Indiana 43017-6221

mkagita@iu.edu

## ABSTRACT

Face recognition is an exciting and emerging field of computer vision with many applications to hardware and devices. Using embedded platforms like the Raspberry Pi, a camera module and open source computer vision libraries like OpenCV, purpose is to add face recognition to Robot car and also facial recognition using free developer version of Kairos Facial Recognition software. In today's modern world, face recognition playing an important role for the purpose of security and surveillance and hence there is a need for an efficient and cost effective system. So the main goal is to explore the feasibility of implementing Raspberry Pi based facial recognition system using conventional face detection and recognition techniques such as Haarcascade detection and Kairos. An obstacle avoidance Robot car is integrated with Raspberry Pi and a camera module aiming at taking face recognition to a level in which the system can identify the humans who are stuck in buildings during earth quakes. Raspberry Pi kit provides the system cost effective and easy to use, with high performance.

## KEYWORDS

Raspberry Pi, Robot Car, Face Recognition, Face Identification, I523, HID319

## 1 INTRODUCTION

An area of application of Computer Vision, one that has always fascinated people, concerns the capability of robots and computers in general to determine, recognize and interact with human counterparts. In this article we will take advantage of the availability of cheap tools for computing and image acquisition, like Raspberry Pi and his dedicated video camera, Camera Pi, and of open source software products for image acquisition and processing, such as OpenCV and SimpleCV, that allow a high level approach to this discipline, and therefore quite a simplified one.

The possibility to locate, within the context of pictures, human beings or their parts like faces, eyes, nose, and so on. This functionality is available in the most advanced photo gallery applications, and it is currently in the implementation phase as for social network applications. Once photos are loaded, the system will scan them to search for people's faces, will find them out and will give a chance to associate a name. If, by chance, the same person is present in different pictures, he/she is recognized and automatically "registered", notwithstanding privacy concerns. This last functionality is the one we previously cited as the one for identification or recognition.

The information age is quickly revolutionizing the way transactions are completed. There is a need for a faster and accurate user identification and authentication method. Face recognition has become one of the most important user identification methods.

Literature survey statistics shows that research work in face recognition system is in its booming era, and in the past forty years, the research in this field has increased exponentially.

The creation of facial biometric data was the first step in creating complete recognition. The second step was being able to match the data to a database of biometrics and associate it with an individual's identity. The human eye can quickly process facial characteristics, but the human brain can store only a few hundred faces reliably. After a while, we tend to forget people's names and often need to relearn information about people. Computers, on the other hand, excel at storing and matching data. Facial recognition software has evolved to the point where computers can process an image and match it against a database of millions of people in seconds.

Law enforcement has led the way with the development of facial recognition systems that can identify criminals against a watch list in real-time. If you've traveled through an airport recently, chances are, your facial biometric data has been captured and matched against a watch list.

Having your biometric data stored in a database has raised privacy concerns. Storing biometric data without consent has been a topic of discussion and debate for privacy groups for years and, in some cases, had led to the creation of policies to protect a person's identity. In addition to privacy concerns, there are also fraud concerns. Having your facial fingerprint matched to what is known as metadata (name, address, and Social Security number, for example) is a major identity theft risk. In order to combat this risk, software vendors have created biometric encryption algorithms to encrypt the data within the database and also provide an almost unbreakable link between the biometric data and the metadata.

## 2 FACE DETECTION

The definition of face detection refers to computer technology that is able to identify the presence of people's faces within digital images. In order to work, face detection applications use machine learning and formulas known as algorithms to detecting human faces within larger images. These larger images might contain numerous objects that aren't faces such as landscapes, buildings and other parts of humans (e.g. legs, shoulders and arms).

Face detection is a broader term than face recognition. Face detection just means that a system is able to identify that there is a human face present in an image or video. Face detection has several applications, only one of which is facial recognition. Face detection can also be used to auto focus cameras. And it can be used to count how many people have entered a particular area. It can even be used for marketing purposes. For example, advertisements can be displayed the moment a face is recognized.

## 2.1 How Face Detection Works

While the process is somewhat complex, face detection algorithms often begin by searching for human eyes. Eyes constitute what is known as a valley region and are one of the easiest features to detect. Once eyes are detected, the algorithm might then attempt to detect facial regions including eyebrows, the mouth, nose, nostrils and the iris. Once the algorithm surmises that it has detected a facial region, it can then apply additional tests to validate whether it has, in fact, detected a face.

## 3 FACE RECOGNITION

Like all biometrics solutions, face recognition technology measures and matches the unique characteristics for the purposes of identification or authentication. Often leveraging a digital or connected camera, facial recognition software can detect faces in images, quantify their features, and then match them against stored templates in a database. Face scanning biometric tech is incredibly versatile and this is reflected in its wide range of potential applications. Face biometrics have the potential to be integrated anywhere you can find a modern camera. Law enforcement agencies the world over use biometric software to scan faces in CCTV footage, as well as to identify persons of interest in the field. Border control deployments use face recognition to verify the identities of travelers. It even has consumer applications. One of the most important applications of face detection, however, is facial recognition. Face recognition describes a biometric technology that goes way beyond recognizing when a human face is present. It actually attempts to establish whose face it is. The process works using a computer application that captures a digital image of an individual's face (sometimes taken from a video frame) and compares it to images in a database of stored records. While facial recognition isn't 100% accurate, it can very accurately determine when there is a strong chance that a person's face matches someone in the database. In the model employed for the extraction of the features from the images, the reference matrices have different shapes, such as the ones that can be seen in figure, that are more suitable for determining the shapes belonging to the human body, like the eyes or the nose. From this comes their denomination of Haar Features, to distinguish them from their original meaning. The same picture shows the shape of the features used by OpenCV and SimpleCV. The presence or not of a Haar feature in a portion of the picture happens by subtracting the median pixel value that are present in the black mask portion, from the median value of the pixels that are present in the clear part of the mask. If the difference is above a certain threshold value, the feature is considered as present. The threshold value is determined, for each feature, during the function training, to detect particular objects or parts of the human body. The learning process materializes itself when presenting to the Vision System the highest possible number of images concerning the objects family that we want to identify, and the highest possible number of images that have nothing to share with the object itself. From the amount of data that are studied, the threshold values are calculated, for each of the features that, in the case of OpenCV and SimpleCV, are memorized as a file in .xml format.

## 4 SOFTWARE AND HARDWARE SPECIFICATIONS

In this project we are using OpenCv in Raspberry pi. This project is used to detect the human Face with the help of OpenCv tool. In order to do object detection with cascade files, you first need cascade files. For the extremely popular tasks, these file already exist.

### 4.1 Software Used

4.1.1 *Raspian OS*. This is the recommended OS for Raspberry Pi. You can also install other OS from third party. Raspbian OS is debian based OS. We can install it from noobs installer.

4.1.2 *Putty*. PuTTY is an SSH and telnet client, developed originally by Simon Tatham for the Windows platform. PuTTY is open source software that is available with source code and is developed and supported by a group of volunteers. Here we are using putty for accessing our raspberry pi remotely.

4.1.3 *OpenCV*. OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. Being a BSD-licensed product, OpenCV makes it easy for businesses to utilize and modify the code. The library has more than 2500 optimized algorithms, which includes a comprehensive set of both classic and state-of-the-art computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects and extract 3D models of objects.

4.1.4 *Python 2 IDE*. Python 2.7.x version Integrated Development Environment is used to compile python program in Raspberry Pi. IDE is a text editor plus terminal combination which is used to work on large projects with complex code bases.

4.1.5 *Kairos Facial Recognition Software*. Kairos is an artificial intelligence company specializing in face recognition. Through computer vision and machine learning, Kairos can recognize faces in videos, photos, and the real-world - making it easier than ever to transform the way your business interacts with people.

- Identity
- Emotions
- Demographics

Kairos navigates the complexities of face analysis technology, so you don't have to. We offer APIs and SDKs any developer can integrate with ease.

### 4.2 Hardware Used

4.2.1 *Raspberry Pi 3*. This is the latest version of raspberry pi. In this we have inbuilt Bluetooth and wi-fi, unlike previously we have to use Wi-Fi dongle in one of its USB port. There are total 40 pins in RPI3. Of the 40 pins, 26 are GPIO pins and the others are power or ground pins (plus two ID EEPROM pins.) There are 4 USB Port and 1 Ethernet slot, one HDMI port, 1 audio output port and 1 micro USB port and also many other things you can see the diagram on right side. And also we have one micro SD card slot wherein we

have to installed the recommended Operating system on micro sd card. There are two ways to interact with your raspberry pi. Either you can interact directly through HDMI port by connecting HDMI to VGA cable, and keyboard and mouse or else you can interact from any system through SSH(Secure Shell)

**4.2.2 Raspberry Pi Camera.** The Raspberry Pi camera module can be used to take high-definition video, as well as stills photographs. It's easy to use for beginners, but has plenty to offer advanced users if you're looking to expand your knowledge. There are lots of examples online of people using it for time-lapse, slow-motion and other video cleverness. You can also use the libraries we bundle with the camera to create effects.

**4.2.3 Robot Car Chassis Kit.** The Mechanical design of the Robot car includes hardware such as motor and wheel placement and body setup. Robot car uses two gear-motors attached to wheels and one free wheel for forward, backward, left and right movements. Free wheel ball is placed at rear side of the robot which helps for 360 degrees free movement [1]. L298N DC Stepper Motor Drive controller is used to control the speed and direction of the two gear motor wheels. Ultrasonic sensors are placed at front side of the robot which is capable to detect the objects on its path.

## 5 SYSTEM ARCHITECTURE

System Architecture consists of following blocks :

- Raspberry Pi
- Raspberry Pi Camera module
- 3 wheel Robot Car kit
- L298N DC Stepper Motor Drive Controller
- 12v and 5v DC batteries

The Mechanical design of the Robot car includes hardware such as motor and wheel placement and body setup. Robot car uses two gear-motors attached to wheels and one free wheel for forward, backward, left and right movements. Free wheel ball is placed at rear side of the robot which helps for 360 degrees free movement. L298N DC Stepper Motor Drive controller is used to control the speed and direction of the two gear motor wheels. Ultrasonic sensors are placed at front side of the robot which is capable to detect the objects on its path. Raspberry Pi Camera module is used to monitor the live stream and recognize the face if its detected.

## 6 SETUP

### 6.1 Connect Raspberry Pi

This section includes connectivity of Raspberry Pi over Wifi.

- Download Raspbian OS to an SD card with a minimum capacity of 8GB.
- Plug in USB power cable, keyboard, mouse and monitor cables to Raspberry Pi.
- Insert the SD card with Raspbian OS into Pi and boot the system. Once the Pi is booted up, a window will appear with Raspbian operating system. Click on Raspbian and Install.
- When the install process has completed, the Raspberry Pi configuration menu (raspi-config) will load. Here set the time and date for your region.

- Enable wifi on upper right corner and connect to wifi sid.

### 6.2 Connect Raspberry Pi Camera Module

- Install the Raspberry Pi Camera module by inserting the cable into the Raspberry Pi.
- The cable slots into the connector situated between the Ethernet and HDMI ports, with the silver connectors facing the HDMI port.
- Boot up your Raspberry Pi and run below commands in command prompt.
- sudo apt-get install python-pip
- sudo apt-get install python-dev
- sudo pip install picamera
- sudo pip install rpio
- From the prompt, run "sudo raspi-config".
- If the "camera" option is not listed, you will need to run a few commands to update your Raspberry Pi. Run "sudo apt-get update" and "sudo apt-get upgrade"

**6.2.1 Enable Camera.** For Face Detection, PiCamera should be enable from Raspberry Pi. Below list of figures shows the detailed steps on how to enable PiCamera from Raspberry Pi.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

### 6.3 Install OpenCV and Required Libraries

OpenCV computer vision library is used to perform face detection and recognition. For this, first need to install OpenCV dependencies on Raspberry Pi. Below commands needs to be executed.

- sudo apt-get update
- sudo apt-get upgrade
- sudo apt-get install build-essential
- cmake pkg-config python-dev libgtk2.0-dev libgtk2.0-zlib1g-dev libpng-dev libjpeg-dev libtiff-dev libjasper-dev libavcodec-dev swig unzip
- Select yes for all options and wait for the libraries and dependencies to be installed

Download opencv-2.4.9 zip file to Raspberry Pi. Change the directory and execute cmake command as below.

- cd opencv-2.4.9
- sudo apt-get install build-essential cmake pkg-config
- sudo apt-get install libjpeg-dev libtiff5-dev libjasper-dev libpng12-dev
- sudo apt-get install python-dev python-numpy libtbb2 libtbb-dev libjpeg-dev libpng-dev libtiff-dev libjasper-dev libdc1394-22-dev
- sudo apt-get install python-opencv
- sudo apt-get install python-matplotlib
- Latest version of OpenCV is now installed in Raspberry Pi

### 6.4 Integration of Raspberry Pi with Robot Car

Raspberry Pi connected with PiCamera is integrated with Robot car to navigate using webserver. During the navigation, robot car will look for human faces using PiCamera and then detects the face. Once the face is detected, python program will call Kairos facial

detection software to identify the person and greet with the name. If the human face is unidentified then robot car will ask human to register their name.

As shown in the figure below, connect a Robot car chassis to raspberry pi and follow the circuit connections.

[Figure 4 about here.]

- Motor1A : 16 (GPIO 23 - Pin 16)
- Motor1B : 18 (GPIO 24 - Pin 18)
- Motor1Enable : 22 (GPIO 25 - Pin 22)
- Motor2A : 21 (GPIO 9 - Pin 21)
- Motor2B : 19 (GPIO 10 - Pin 19)
- Motor2Enable : 23 (GPIO 11 - Pin 23)

## 7 CODE EXPLANATION

### 7.1 Face Detection

---

```
from picamera.array import PiRGBArray
from picamera import PiCamera
import time
import cv2
import sys
import imutils
from fractions import Fraction
import base64
import requests
import json
import random
import os
```

---

```
# Get user supplied values
cascPath = './haarcascade_frontalface_default.xml'
```

---

```
# Create the haar cascade
faceCascade = cv2.CascadeClassifier(cascPath)
```

---

```
# initialize the camera and grab a reference to the raw
# camera capture
camera = PiCamera()
camera.resolution = (160, 120)
camera framerate = 32
rawCapture = PiRGBArray(camera, size=(160, 120))
```

---

```
# allow the camera to warmup
time.sleep(0.1)
lastTime = time.time()*1000.0
# capture frames from the camera
for frame in camera.capture_continuous(rawCapture,
    format="bgr", use_video_port=True):
    # grab the raw NumPy array representing the image, then
    # initialize the timestamp
    # and occupied/unoccupied text
    image = frame.array
    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

    # Detect faces in the image
    faces = faceCascade.detectMultiScale(
```

```
gray,
scaleFactor=1.1,
minNeighbors=5,
minSize=(30, 30),
flags = cv2.cv.CV_HAAR_SCALE_IMAGE
)
print time.time()*1000.0-lastTime, " Found {0}
faces!".format(len(faces))
lastTime = time.time()*1000.0

# Draw a rectangle around the faces
for (x, y, w, h) in faces:
    cv2.rectangle(image, (x+w/2, y+h/2), int((w+h)/3),
(255, 255, 255), 1)
# show the frame
cv2.imshow("Frame", image)
key = cv2.waitKey(1) & 0xFF
if len(faces) == 1:
    print("Taking image...")
camera.capture("foo.jpg")
os.system('espeak "Human face detected"')
inputImage= "./foo.jpg"
del camera
break
# clear the stream in preparation for the next frame
rawCapture.truncate(0)

# if the `q` key was pressed, break from the loop
if key == ord("q"):
    del camera
    exit()
```

---

### 7.2 Face Recognition

---

```
KAIROS = "api.kairos"
KairosGallery = 'MyFace'
KairosConfig = './kairos_config.json'
```

---

```
def trainKairos(image, name):
    global KairosGallery
    headers = {
        'app_id': 'd39fc1b1',
        'app_key': '468d508d9463c8d24395926adabb1769'
    }
    data = {
        'image': base64.b64encode(image),
        'gallery_name': KairosGallery,
        'subject_id': name
    }
    r = requests.post('http://api.kairos.com/enroll',
                      headers=headers, data=json.dumps(data))
    print(r.text)
    return(None)
```

---

```
class Recognize():
    def __init__(self, API, config_file):
        self.api = API
        self.config = config_file
```

```

#define recognize(self, image_path):
#    return self._recognizeKairos(image_path)

def recognizeKairos(self, image):
    with open(image, "rb") as image_file:
        encoded_string =
            base64.b64encode(image_file.read())
    with open(self.config, "rb") as config_file:
        config = json.loads(config_file.read())
    data = {
        "image": encoded_string,
        "gallery_name": config["gallery_name"]
    }

    headers = {
        "Content-Type": "application/json",
        "app_id": config["app_id"],
        "app_key": config["app_key"]
    }
    try:
        r =
            requests.post("https://api.kairos.com/recognize",
                          headers=headers, data=json.dumps(data))
        data = r.json()
    print data
#    print json.dumps(data, indent=4)
    faces = []
    if "images" in data:
        for obj in data["images"]:
            if obj["transaction"]["status"] == "success":
                face_obj = {}
                face_obj["person"] =
                    obj["transaction"]["subject_id"]
                .decode("utf_8")
                #face_obj["faceid"] =
                #    obj["candidates"][0]["face_id"]
                #.decode("utf_8")
                face_obj["confidence"] =
                    obj["transaction"]["confidence"]
                faces.append(face_obj)
            elif obj["transaction"]["status"] == "failure":
                face_obj = {}
                face_obj["person"] = "unidentified"
                face_obj["confidence"] = 0
                faces.append(face_obj)
            else:
                print "its in last loop"
    return faces
    except requests.exceptions.RequestException as
        exception:
            print exception
    return None

```

---

```

if __name__ == "__main__":
    r = Recognize(KAIROS, "kairos_config.json")
    x = r.recognizeKairos(inputImage)

    #print x
    #print x["person"]
    #print x[0]["person"]

```

```

string1 = x[0]["person"]
#print string1
os.system('espeak "Hello...""{}''.format(string1))
if x[0]["person"] == "unidentified":
    os.system('espeak "Please enter your name to
    Register"')
    nameToRegister = raw_input("Please enter your name
        to Register :")
    binaryData = open(inputImage, 'rb').read()
    print('Enrolling to Kairos')
    trainKairos(binaryData, nameToRegister)
    print "You are now Registered as :", nameToRegister
    os.system('espeak
        "Hello...""{}''.format(nameToRegister))
    exit()

```

---

### 7.3 Robot Car Navigation

---

```

import RPi.GPIO as GPIO
from time import sleep

```

---

```

GPIO.setmode(GPIO.BRD)

```

---

```

#Connecting two wheel motors to Raspberry Pi GPIO
#Left Motor (Motor 1) connections
Motor1A = 16 #(GPIO 23 - Pin 16)
Motor1B = 18 #(GPIO 24 - Pin 18)
Motor1Enable = 22 #(GPIO 25 - Pin 22)

```

```

#Right Motor (Motor 2) Connections
Motor2A = 21 #(GPIO 9 - Pin 21)
Motor2B = 19 #(GPIO 10 - Pin 19)
Motor2Enable = 23 #(GPIO 11 - Pin 23)

```

---

```

#Output of Motors to set as OUT
GPIO.setup(Motor1A,GPIO.OUT)
GPIO.setup(Motor1B,GPIO.OUT)
GPIO.setup(Motor1Enable,GPIO.OUT)
GPIO.setup(Motor2A,GPIO.OUT)
GPIO.setup(Motor2B,GPIO.OUT)
GPIO.setup(Motor2Enable,GPIO.OUT)

```

---

```

# Defining function for Robot car to move forward
def forward():
    GPIO.output(Motor1A,GPIO.HIGH)
    GPIO.output(Motor1B,GPIO.LOW)
    GPIO.output(Motor1Enable,GPIO.HIGH)
    GPIO.output(Motor2A,GPIO.HIGH)
    GPIO.output(Motor2B,GPIO.LOW)
    GPIO.output(Motor2Enable,GPIO.HIGH)

    sleep(2)

```

---

```

# Defining function for Robot car to move backward
def backward():
    GPIO.output(Motor1A,GPIO.LOW)

```

```

GPIO.output(Motor1B,GPIO.HIGH)
GPIO.output(Motor1Enable,GPIO.HIGH)
GPIO.output(Motor2A,GPIO.LOW)
GPIO.output(Motor2B,GPIO.HIGH)
GPIO.output(Motor2Enable,GPIO.HIGH)

sleep(2)

# Defining function for Robot car to turn right
def turnRight():
    print("Going Right")
    GPIO.output(Motor1A,GPIO.HIGH)
    GPIO.output(Motor1B,GPIO.LOW)
    GPIO.output(Motor1Enable,GPIO.HIGH)
    GPIO.output(Motor2A,GPIO.LOW)
    GPIO.output(Motor2B,GPIO.LOW)
    GPIO.output(Motor2Enable,GPIO.LOW)

    sleep(2)

# Defining function for Robot car to turn left
def turnLeft():
    print("Going Left")
    GPIO.output(Motor1A,GPIO.LOW)
    GPIO.output(Motor1B,GPIO.LOW)
    GPIO.output(Motor1Enable,GPIO.LOW)
    GPIO.output(Motor2A,GPIO.HIGH)
    GPIO.output(Motor2B,GPIO.LOW)
    GPIO.output(Motor2Enable,GPIO.HIGH)

    sleep(2)

# Defining function for Robot car to stop
def stop():
    print("Stopping")
    GPIO.output(Motor1A,GPIO.LOW)
    GPIO.output(Motor1B,GPIO.LOW)
    GPIO.output(Motor1Enable,GPIO.LOW)
    GPIO.output(Motor2A,GPIO.LOW)
    GPIO.output(Motor2B,GPIO.LOW)
    GPIO.output(Motor2Enable,GPIO.LOW)

```

---

## 7.4 Controlling Robot Car using webserver

```

from flask import Flask, render_template, request,
               redirect, url_for, make_response
import RPi.GPIO as GPIO
import motors

#set up GPIO
GPIO.setmode(GPIO.BOARD)

#set up flask server
app = Flask(__name__)

#when the root IP is selected, return index.html page
@app.route('/')

```

```

def index():
    return render_template('index.html')

#receive which pin to change from the button press on
#index.html
#each button returns a number that triggers a command in
>this function
#
#Uses methods from motors.py to send commands to the GPIO
#to operate the motors
@app.route('/<changePin>', methods=['POST'])
def reroute(changePin):
    changePin = int(changePin) #cast changePin to an int

    if changePin == 1:
        motors.turnLeft()
    elif changePin == 2:
        motors.forward()
    elif changePin == 3:
        motors.turnRight()
    elif changePin == 4:
        motors.backward()
    else:
        motors.stop()

    response = make_response(redirect(url_for('index')))
    return(response)

#set up the server in debug mode to the port 8000
app.run(debug=True, host='0.0.0.0', port=8000)

```

---

## 8 APPLICATION

There are lots of applications of face recognition. Face recognition is already being used to unlock phones and specific applications. Face recognition is also used for biometric surveillance. Banks, retail stores, stadiums, airports and other facilities use facial recognition to reduce crime and prevent violence.

## 9 CONCLUSION

### ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions in writing this paper.

### REFERENCES

- [1] Arduino. 2015. Arduino Software (IDE). (2015). <https://www.arduino.cc/en/Guide/Environment>

#### LIST OF FIGURES

1	Edit raspi-config file from command line	8
2	Select Camera from the options	8
3	Enable Camera	9
4	Raspberry Pi Robot Car Integration	10

```
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.
```

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Mon May 29 18:17:10 2017
pi@raspberrypi:~ $ sudo raspi-config
```

Figure 1: Edit raspi-config file from command line

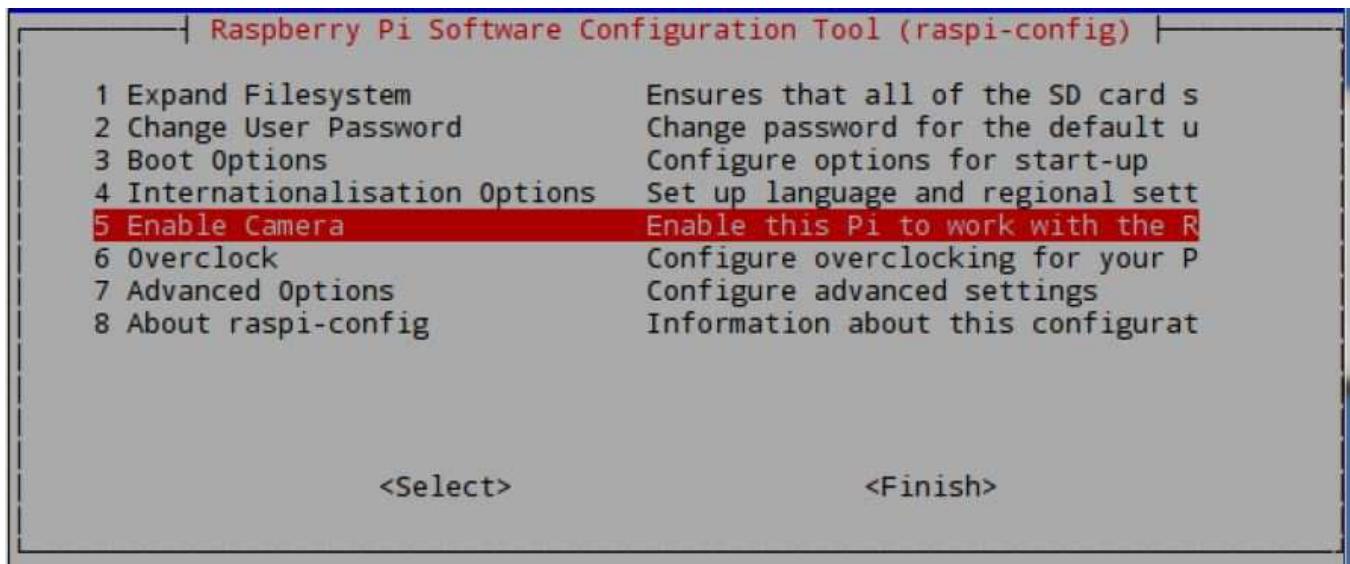


Figure 2: Select Camera from the options

Would you like the camera interface to be enabled?

<Yes>

<No>

Figure 3: Enable Camera

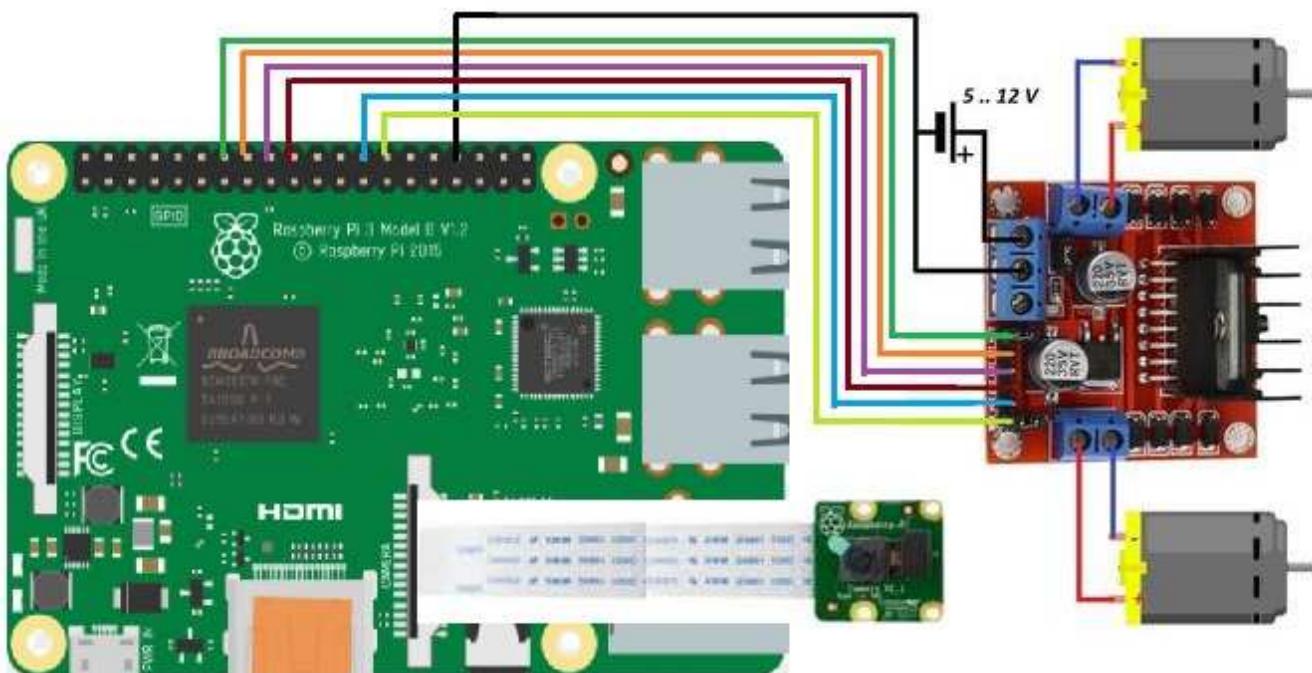


Figure 4: Raspberry Pi Robot Car Integration

## 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

## Compliance Report

```
=====
name: Mani Kumar Kagita
hid: 319
paper1: 100%
paper2: 100%
project: 80%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
10
wc 319 project 10 3779 report.tex
wc 319 project 10 3664 report.pdf
wc 319 project 10 24 report.bib
```

```
find "
```

---

```
253: for frame in camera.capture_continuous(rawCapture, format="bgr",
    use_video_port=True):
267: print time.time()*1000.0-lastTime," Found {0}
    faces!".format(len(faces))
274: cv2.imshow("Frame", image)
277: print("Taking image...")
278: camera.capture("foo.jpg")
279: os.system('espeak "Human face detected"')
280: inputImage= "./foo.jpg"
287: if key == ord("q"):
295: KAIROS = "api.kairos"
327: with open(image, "rb") as image_file:
```

```
329: with open(self.config, "rb") as config_file:  
330:  
331:     config = json.load(config_file)  
332:     "image": encoded_string,  
333:     "gallery_name": config["gallery_name"]  
334:  
335:     "Content-Type": "application/json",  
336:  
337:     "app_id": config["app_id"],  
338:     "app_key": config["app_key"]  
339:  
340: r = requests.post("https://api.kairos.com/recognize",  
341:                      headers=headers, data=json.dumps(data))  
342:  
343: if "images" in data:  
344:  
345:     for obj in data["images"]:  
346:  
347:         if obj["transaction"]["status"] == "success":  
348:  
349:             face_obj["person"] = obj["transaction"]["subject_id"]  
350:  
351:             face_obj["faceid"] = obj["candidates"][0]["face_id"]  
352:  
353:             .decode("utf_8")  
354:  
355:             face_obj["confidence"] = obj["transaction"]["confidence"]  
356:  
357:         elif obj["transaction"]["status"] == "failure":  
358:  
359:             face_obj["person"] = "unidentified"  
360:  
361:             face_obj["confidence"] = 0  
362:  
363: print "its in last loop"  
364:  
365: if __name__ == "__main__":  
366:  
367:     r = Recognize(KAIROS, "kairos_config.json")  
368:  
369:     #print x["person"]  
370:  
371:     #print x[0]["person"]
```

```
378: string1 = x[0] ["person"]

380: os.system('espeak "Hello...""{}"'.format(string1))

381: if x[0] ["person"] == "unidentified":

382: os.system('espeak "Please enter your name to Register"')

383: nameToRegister = raw_input("Please enter your name to Register
   :")

387: print "You are now Registered as :", nameToRegister

388: os.system('espeak "Hello...""{}"'.format(nameToRegister))

453: print("Going Right")

467: print("Going Left")

481: print("Stopping")

passed: False

find footnote
-----
passed: True

find input{format/i523}
-----
4: \input{format/i523}

passed: True

find input{format/final}
-----
passed: False

floats
-----
160: \begin{figure}[ht!]
161: \includegraphics[width=\columnwidth]{images/enablecamera1.jpg}
```

```
165: \begin{figure}[ht!]
166: \includegraphics[width=\columnwidth]{images/enablecamera2.jpg}
170: \begin{figure}[ht!]
171: \includegraphics[width=\columnwidth]{images/enablecamera3.jpg}
200: \begin{figure}[ht!]
201: \includegraphics[width=\columnwidth]{images/RaspPi_Robot.jpg}
```

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

```
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: table and above may be used improperly
```

```
72: In the model employed for the extraction of the features from the
    images, the reference matrices have different shapes, such as the
    ones that can be seen in figure, that are more suitable for
    determining the shapes belonging to the human body, like the eyes
    or the nose. From this comes their denomination of Haar Features,
    to distinguish them from their original meaning. The same picture
    shows the shape of the features used by OpenCV and SimpleCV. The
    presence or not of a Haar feature in a portion of the picture
```

happens by subtracting the median pixel value that are present in the black mask portion, from the median value of the pixels that are present in the clear part of the mask. If the difference is above a certain threshold value, the feature is considered as present. The threshold value is determined, for each feature, during the function training, to detect particular objects or parts of the human body. The learning process materializes itself when presenting to the Vision System the highest possible number of images concerning the objects family that we want to identify, and the highest possible number of images that have nothing to share with the object itself. From the amount of data that are studied, the threshold values are calculated, for each of the features that, in the case of OpenCV and SimpleCV, are memorized as a file in .xml format.

WARNING: figure and above may be used improperly

72: In the model employed for the extraction of the features from the images, the reference matrices have different shapes, such as the ones that can be seen in figure, that are more suitable for determining the shapes belonging to the human body, like the eyes or the nose. From this comes their denomination of Haar Features, to distinguish them from their original meaning. The same picture shows the shape of the features used by OpenCV and SimpleCV. The presence or not of a Haar feature in a portion of the picture happens by subtracting the median pixel value that are present in the black mask portion, from the median value of the pixels that are present in the clear part of the mask. If the difference is above a certain threshold value, the feature is considered as present. The threshold value is determined, for each feature, during the function training, to detect particular objects or parts of the human body. The learning process materializes itself when presenting to the Vision System the highest possible number of images concerning the objects family that we want to identify, and the highest possible number of images that have nothing to share with the object itself. From the amount of data that are studied, the threshold values are calculated, for each of the features that, in the case of OpenCV and SimpleCV, are memorized as a file in .xml format.

WARNING: figure and below may be used improperly

158: For Face Detection, PiCamera should be enable from Raspberry Pi. Below list of figures shows the detailed steps on how to enable PiCamera from Raspberry Pi.

WARNING: figure and below may be used improperly

198: As shown in the figure below, connect a Robot car chassis to raspberry pi and follow the circuit connections.

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 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

# The Intersection of Big Data and IoT

Peter Russell  
Indiana University  
petrusse@iu.edu

## ABSTRACT

Big Data and IoT share a symbiotic relationship with one another that is leading to incredible innovations that were inconceivable just 18 years ago. As a result of this relationship, it has become easier than ever for individuals to customize and monitor various elements of their life if they choose to do so. A project is undertaken to demonstrate how accessible IoT has become to leverage Big Data analysis, how IoT and Big Data are being utilized together in some of the most interesting current use cases and how the technology will need to adapt in coming years.

## KEYWORDS

i523, HID 334, Edge Computing, Raspberry Pi, IoT

## 1 INTRODUCTION

In 2020, it is estimated that 95 percent of electronics will contain IoT technology [41]. This technology, commonly abbreviated for the “Internet of Things”, is expected to be so pervasive due to how the technology is defined and how impactful it has already become.

At the highest level, IoT is intended to describe devices that collect and relay information via the Internet. This leaves IoT is broadly defined in the type of application, which could come in the form of a phone, vehicle, a home device like a thermostat or television, but the technology is rather specific in its intended application. That is, these devices are generally made to serve a single purpose and they are extremely adept at that function. In its most powerful applications, massive data sets are created from these individual IoT devices as they are synthesized together to meet a larger need, as we will see.

The specific purpose of an IoT device is what differentiates this emerging technology from traditional computers. With the exception of recent developments, which will be explored in depth, early IoT devices were not intended to do the heavy computing like a computer would do. However, with rapid advancements in computing power and speed, the line differentiating the two has begun to blur. It is this increase in computing power that has lead IoT and Big Data to have a cooperative relationship to create some of the most exciting technology that's available today.

The ability to collect and process more data has increased the utility of these devices as they're able to become more personalized, spurring tremendous growth recently, on the order of 30% annually. In 2017, it is expected to be the year that the number of IoT devices exceeds the number of people on Earth [18]. This personalization is not without consequence though, which will be discussed later, so the relationship between IoT and Big Data is still evolving.

To begin, we examine how the IoT came to be and continuously evaluate how it is integrated with Big Data. Then, a demonstrative project will be outlined to show the Big Data can be used in an IoT

device. Next, we discuss high level implementations of these technologies in modern use before discussing some of the challenges the industry is facing.

## 2 EMERGENCE OF IOT

Given the massive and recent popularity of IoT, it might be a surprise to some to learn that this concept has been around since 1999. The idea to have multiple, remote devices communicating with one another to gain insights to a single problems was originally conceived by Kevin Ashton as a solution to supply chain management [16]. At that point, the idea was ahead of its time as the internet was still gaining widespread adoption. However, as computing power and sensor costs have declined, IoT has become a main an indispensable aspect of most people's lives. One such example could be the integration of global positioning systems (GPS) into cellphones, which was introduced in just 2004, but has become a staple for nearly every phone released [50].

### 2.1 What Defines IoT?

Given the ascension of so many new technologies, it could be helpful to understand what technically constitutes an IoT device. This will be useful later when discussing the sample project undertaken and how these both relate to Big Data analysis.

As stated earlier, at a high level, IoT is meant to describe devices that use internal or external sensors to connect to the Internet. These sensors could come in the form of the well known types, such as Wi-Fi, Bluetooth or RFID, to the perhaps less widely known, such as NFC (Near Field Communication) or Zigbee [43].

For these IoT devices, the Internet allows them to be tremendously influential in the advancement of Big Data by virtue of the amount of data the devices are able to collect. Specifically, IoT allows its users to quantify the world around them in ways that were previously not possible. For corporations, this yields tremendous advantages when it comes to business planning or equipment monitoring. For example, the average wind farm can generate 150,000 data points *per second* and an engine turbine could give 500 gigabytes of data [28]. Additionally, for individuals, IoT enables people to monitor their activity on a daily basis through wearable fitness devices and customize their homes to save on energy consumption. It has been estimated the average household generates approximately 2,000 gigabytes of data a year and this is expected to increase five fold by 2020 [52]. As we will explore, this rapid increase is due in large part to the computing power of the individual devices, which allow for a greater volume of data to be collected. For example, if a person enjoys a simple bike ride and purchases the Garmin Edge 500 watch, on a single ride they are producing data across 61 different variables for statistics such as heart rate, elevation gained, cadence and output produced continuously for the duration of their ride [17].

## 3 IOT PROJECT

### 3.1 IoT Device

The Raspberry Pi 3 (Model B) was chosen as an example IoT device to demonstrate how these devices can be used to from Big Data analytics. The Raspberry Pi has drawn tremendous accolades for its initiative to get inexpensive, but powerful computing power into the hands of aspiring programmers and hobbyists. Equipped with 1GB of RAM, a 1.2GHz quad-core processor and Bluetooth/Wi-Fi capability, one can purchase the device for just \$35.

### 3.2 Description

The goal of this device is to create a personalized interface that gives the user a morning snapshot for relevant, important information to begin their day. As it relates to IoT, this project uses IoT technology through Wi-Fi to source the output of Big Data projects undertaken by others (ie. Google and Weather Underground as will be shown).

### 3.3 Implementation

For those unfamiliar with the Raspberry Pi, the initial setup could be somewhat intimidating the first time around. Specifically, the Raspberry Pi comes as a truly blank slate and to begin using it, one will need to write the OS, Raspbian, onto the Pi. Several tutorials are available online to get to the desktop, so in the interest of brevity, the discussion below will assume that the user has been able to successfully get the Pi operational and to the Linux prompt with Python installed to run the script.

The application was developed using Python, utilizing the Kivy package for GUI development, the requests and Beautiful Soup packages for the user location, news stories and sports scores along with Yahoo Weather/Weather Underground via the Weather package. Outside of these, standard Python library packages were used.

### 3.4 Results

As part of the display, a continuous running clock was added, which necessitated the application to be on a constant refresh. This was implemented successfully and at a relatively low cost with no significant delays. The total build of the application consumed 966,565 bytes with each refresh using just 1032 bytes. At the initialization of the program, the application uses the user's IP address to find their zip code to populate the local weather forecast and local news. For the weather, the user will get the current temperature with a high/low for the current day and each day in the five day forecast.

Additionally, as news stories are published to the WSJ news feed, they will be read into the application and refreshed. Stories are shown in chronological order along with their time stamp of publication and each is hyperlinked directly to the full story if the user wants more information.

### 3.5 Application to Big Data

One of the main benefits of IoT is synthesizing data across numerous platforms and data sources into a desired output. Our desired output is a one-stop interface with interesting information that can be displayed anywhere with an outlet, internet connection and monitor.

The various components of the monitor are only made possible by the creativity by the providers in solving difficult Big Data issues, such as the clustering of news (Google) and weather forecasting (Yahoo/Weather Underground). Each provider's relationship to Big Data is worth examination and will be the topic of the following sections.

### 3.6 Google News

*3.6.1 Big Data Description.* Google News has evolved into a central source of information for how a large share of the population receives its news. In fact, as a display of the trust that users place in Google to deliver the most information in the most efficient manner, it was found that users are more likely to trust a Google news headline than that same headline from the original source [26]. Additionally, 44% of users were found to read nothing more than the headlines [22]. This is a testament to their ability to simplify the universe of world news into succinct rankings.

Entering into its fifteenth year, Google News aggregates from 50,000 news sources worldwide across 30 different languages. In 2012, they reported the division was receiving 1 billion unique visits a week[4]. For reference, major individual news providers, such as CNN and the New York Times receive 125 million and 99 million unique visitors per month [3]. These statistics further demonstrate Google's successful navigation of the Big Data problem for news stories in the eyes of its users. It's relatively clear why this is an important Big Data problem, but one might be curious how they're able to effectively navigate the problem. Unfortunately, the full design from start to finish is a well kept secret, but pieces have been released and one can piece together a mosaic view of what might be going on under the hood. The decision to not disclose the techniques for ranking news stories is understandable, but it has been a lightening rod of controversy nonetheless. Some view the decision of Google's scoring as effectively acting as a censor for the internet while they maintain it is to keep the integrity of the algorithm so that news stories cannot be written purely for traffic, known as search engine optimization [12].

On the surface, one might question the economic value of Google News to the larger company since it is a free service for both users and providers. However, there is a tremendous amount to be gained in solving this Big Data problem. Even though Google does not show ads on its news site, it was estimated to generate spillover traffic into its search engine that leaves the News entity worth \$100 million in 2008 [60]. The current valuation is undoubtedly higher, but it remains undisclosed. So while there are profits to be made for Google in this quest, publishers of these stories have a tremendous amount of interest in this problem as well. Some providers don't believe content should be indexed to Google's search algorithm for free and Google should pay them for their investigative research. One such provider, who happened to be Germany's largest news source, decided to remove themselves from the index for two weeks. The results were devastating for the site as traffic through the site dropped by 40% [59]. It was a quick lesson in how critical positioning can be in the Big Data world of news aggregation.

*3.6.2 Big Data Solution.* Just as news is constantly evolving, so are Google's solutions to this Big Data problem. As we've seen recently, news aggregation services are under pressure to become

more intelligent on what news is shown in the hopes of preventing fake news from making it into the top results.

The technical specifics of what Google is implemented has largely been kept under wraps, but we did learn a few of the techniques and platforms in 2007. In at least the early versions, Google used MinHash, Probabilistic Latent Semantic Indexing and Covisitation to solve this Big Data. Specifically, these methods will compare historical clicks with other similar users for recommendations, decipher key words and phrases from an article for grouping and track how news stories are clicked within a certain time frame to find stories that were read successively. For processing these queries, Google uses MapReduce and Hadoop architecture [55].

For the inputs into these algorithms, Google will analyze several metrics of a provider to see how they should be ranked along with the user preferences. These metrics include things like how large the staff is, how many articles they put out, how many websites reference that news source (PageRank) and the breadth of news topics covered [15].

### 3.7 Weather Underground via Yahoo API

**3.7.1 Big Data Description.** Weather is a primary concern in business planning for many industries, such as airlines or agriculture. As a result, companies are willing to dedicate a tremendous amount of resources towards accurate forecasts. One of the most innovative companies and a great example of the intersection of IoT and Big Data is Weather Underground.

Weather Underground is a weather forecasting service that was once owned by the Weather Channel and recently, partially by IBM to integrate with its growing IoT ecosystem. What makes the company unique is how their forecasts are formulated. In their model, they couple traditional forecasting tools with IoT. The traditional readings come from the National Weather Service (NWS), which aggregates data from airports and weather balloons. IoT has lead to a new dimension of forecasting as personalized weather stations are distributed to its users for live forecasts in places that traditional instruments might not be available. As of now, they have 250,000 users set up on the platform. This setup provides an additional layer of information, yielding more frequent data, longer forecast windows and greater certainty for a given area. Namely, users can get new forecasts every 15 minutes (versus every 4 hours on the NWS) and forecasts up to two weeks in advance (compared to one week for NWS) [54]. This use of the IoT, specifically edge computing, which will be expanded on later, provides a tremendous example of how IoT can be used to enhance Big Data analysis.

For those that choose to participate in the service, they will purchase a Personal Weather Station (PWS) that allows them to measure temperature, humidity, pressure, rainfall, wind speed and direction via sensors. The major advantage of the PWS comes from its pressure and wind metrics as users can get a better idea of humidity and wind chill, giving a more accurate representation of current conditions. Neither of these are available through the NWS. In the end, this amounts to around 3 billion data points for the Weather Underground model, servicing around 26 billion inquiries a day [24].

**3.7.2 Big Data Solution.** To process its data in the past, which amounts to multiple terabytes daily, Weather Underground has

stored its forecasts, radar data and satellite data using Apache Hadoop and Amazon Web Services [47]. In fact, IBM has stated a large reason for their motivation to have an ownership stake in the company was due to the cloud infrastructure that Weather Underground had built for fielding the massive volume of requests and forecasts it processes daily.

### 3.8 Limitations

While this demonstrative program shows what an IoT device can achieve with Big Data, the most influential uses of IoT come when devices are able to communicate with one another and create more data that can be implemented back into future improvements in the device, as we have seen. The limitation of this current program is that it is a singular instance of the application. Allowing multiple applications to be deployed where information can be collected on *volunteered* information would be orders of magnitude more difficult to implement, but could be an interesting addition to gain better insight on the user base. Then, this data can be pooled together for how the application could be tailored to meet geographic or demographic preferences.

Additionally, with a large enough user base, we would be interested in tracking the number of downloads, active users and which panels of the display are clicked most often. All of these metrics can be readily accessed through integration with Google Analytics, which allows one to analyze different events within the application [20].

## 4 EDGE COMPUTING

With the Personal Weather Systems, we were able to see how IoT can complement Big Data analysis. This is an example of an emerging technology known as “edge computing” that is transforming how the cloud is utilized.

Edge computing gains its name from how the information being processed by the device. Prior to this recent innovation, information was gathered, sent to the cloud, processed there, and then the output is pushed back to the device. Namely, it was a centralized process. However, with edge computing, devices are more intelligent in what information they choose to send, providing a much more efficient process. For example, rather than having a camera monitor an area constantly, even when there is no motion, modern IoT cameras have been equipped with motion detection so information is only sent when there is something to actually record. Since this decision and processing is made on the actual device, it is considered to be at the *edge* of the network.

Traditionally, IoT devices that were intended to work in conjunction, such as surveillance cameras, were simple in their functionality and storage. Namely, a group of cameras would record individually and send their results back to a central server. However, with improvements in image quality, this becomes a Big Data problem very quickly as these cameras are running around the clock collecting footage. In the historical model of a centralized server, this setup eventually creates problems as bandwidth and storage issues emerge. These limitations are the problems that edge computing seeks to circumvent and has become a major catalyst in the growth of IoT devices [51].

Circling back to the original project that was undertaken, the application benefited from edge computing through the weather data, but the device itself serves as a great example of why edge computing is even possible in the first place. Specifically, it is possible due to the dramatic decrease in computing costs. For the cost of \$10 one can get a single-board computer with 1 GHz and 512 MB RAM through the Raspberry Pi. This type of processing is close to becoming the majority as it is expected that by 2019, 45% of all data collected by IoT devices will be processed at the edge of the network [37]. As we will see, this technology is allowing early adopters to gain unique, real-time insights through Big Data analytics into the health and composition of their businesses.

## 4.1 Use Case: Fraud Detection

Fraudulent transactions represent just 1% of all transactions. However, while the relative size of these transactions to the overall market are small, their absolute impact is enormously detrimental to merchants and financial services companies. In 2015, total fraudulent transactions created damages of \$22 billion [29].

The economic impact of these transactions has given these companies a tremendous incentive to innovate their way out of this problem. The marriage of IoT and Big Data has now provided them the opportunity to have near real-time analytics, which is necessary to effectively manage the problem. This is because the approval process for a transaction needs to be as close to instantaneous as possible. If shortcuts are taken in the analysis to increase speed, fraudulent transactions could slip through and not get flagged. IoT has helped make this trade-off between accuracy and speed less of an issue with new innovations, such as Visa's Ready program.

Visa Ready is an innovative program enables payments through IoT for both security and convenience. Instead of traditional means of payment authorization, such as simply swiping your credit card at a vendor, IoT enables Visa to take advantage of improvements in biometric technology [57]. Visa has introduced multi-dimension verification through biometrics by letting users endorse a payment through their fingerprint, iris scan, face scan and even their voice [58]. This type of technology is gaining adoption and there are expected for be 500 million devices with biometric sensors by 2018 and 26 billion by 2020 [48].

Complementing biometric data, as IoT devices become more mainstream, companies such as FICO are using behavioral data in fortifying their analysis of whether a transaction is fraudulent or not. This type of analysis is not new in and of itself as it has been established as a way to identify e-commerce fraud, but the application through IoT is providing a new dimension of analysis. Traditionally, behavior data was tracked to see how a user interacts with a website to reduce the number of false positives that get flagged, which could occur if a user was on a business trip and abruptly logged into their account to buy something from an IP in another country [14]. With IoT, this adds a tremendous amount of data to an already Big Data problem. Now these companies will have data on how users interact with an IoT device, such as how they hold their device in the case of a phone or their tendencies when using the keyboard [25]. From a business perspective, this all occurs in the background without the user's experience without the product being interrupted.

As a testament to the future of this relationship between IoT and Big Data, Visa has partnered with IBM. This was done in an effort to gain maximum benefit from this new biometric technology by leveraging Visa's payment infrastructure with IBM's efforts in artificial intelligence and Big Data analysis with IBM Watson [31].

## 4.2 Use Case: Autonomous Vehicles

In many ways, autonomous vehicles represent the pinnacle of edge computing to date in unifying IoT and Big Data. Among its many goals, this technology is trying to use Big Data to resolve one of the modern tragic realities of our modern world - automobile fatalities. Automobile accidents cause 1.2 million deaths a year, 94% of which are attributable to human error [30]. For this reason, in conjunction with expected energy savings from car designs with this technology, the technology is expected to experience adoption rates that rivals mobile phones with significantly more impact [23]. Traditional car makers have taken notice of the potential future and as an example of this, General Motors recently hired an Uber engineer to lead its self-driving initiative as the company's first ever Chief Technology Officer [5].

The relation of autonomous cars to IoT via edge computing is once again out of necessity for real-time functionality. A car that processes it should stop two seconds too late is as potentially useful as never making the calculation in the first place, so timing is of the utmost importance. Amazing progress has already been made in the speed and complexity of calculations these autonomous vehicles can handle. One of the highest profile graphic card manufacturers, Nvidia, recently announced their system for autonomous vehicles at the rate of 320 *trillion* operations a second [21]. Since these vehicles are equipped with various types of sensors to process its environment, this type of computing power is a near necessity to tackle this Big Data problem in real-time.

Kevin Ashton's original vision for the IoT was to have an accurate view of inventory as RFID scanners synced over the Internet. In just 18 short years, these autonomous vehicles are achieving the same end of communication with one another on an incredible scale. In what's known as "vehicle to vehicle communication" autonomous cars will be able to send one another information on important considerations, such as road hazards or conditions, allowing GPS to take the most optimal route to its destination. Similarly, speed limit signs can take weather conditions into account, dynamically adjust the speed limit of the road and relay this to the car's navigation system [1].

The companies that are pursuing autonomous driving are largely having the cars learn through the experiences of its sensors. It would be impossible to code every possible scenario a car could face, so instead, data is collected from the various sensors and loaded to the cloud for later analysis. For example, Tesla is accumulating a million miles worth of data across its sensors every 10 hours, leaving it with 780 million through mid-2016 [7]. These sensors on board, which will be briefly described to show their application, are expected to generate 4,000 gigabytes of data *daily* [38] [19]. This is another instance of the familiar union between IoT and Big Data.

### 4.3 Use Case: Health Care

The United States, like the world as a whole, is experiencing an aging crisis in its population. In both the world and the United States, the number of adults aged 65 and over is expected to double by 2025. In the United States, this demographic of the population will move from 15% to 25%. While this jump is not negligible, the most alarming aspect of this statistic is that in 2010, the elderly portion of the population was just 10%, but accounted for 34% of medical expenditures [34].

For this reason of high future expenditures, much of today's public policy debates center around how resources will be pooled to meet this not so distant future need. Currently, one of the most promising use cases for edge computing is coming from health care and how the technology can be used to provide better care to a wider range of people.

Through edge computing, doctors have the ability to gain insights into their patients through sensors that can be worn by their patients, such as a heart monitor. This allows for early identification of irregular patterns and allows for an earlier diagnosis, potentially saving the patient's life compared to earlier times when a heart attack could strike abruptly without warning. This usage is directly related to Big Data as doctors now can get continuous, real-time assessments of their patients. This makes way to more accurate future diagnoses as more insights can be gleaned between the true cause and effect of a particular ailment [42].

Outside of data analysis by doctors, the patients themselves are expected to receive numerous benefits from this type of monitoring. Namely, those who are less mobile no longer need to make a physical trip to see the doctor as the doctor has the diagnostics they typically need and at a much more granular level [6].

In addition to the elderly, this type of real-time feedback system through edge computing can be incredibly transformative for those with health conditions that require nearly continuous monitoring. One such example has been demonstrated with epileptic patients. An edge computing solution has been introduced that epileptic patients can use and if a patient experiences an epileptic episode, an immediate alert is sent to family members and doctors [53]. This type of technology is only possible through edge computing because the alerts are triggered by monitoring historical metrics versus live readings in areas like heart rate and sudden movements. The delay that would be incurred by sending this data to the cloud and waiting for a response would have too much latency to be an effective solution to this problem.

Another promising area for edge computing within health care is for those suffering from mental diseases, such as dementia or Alzheimer's. With this technology, family members can monitor and set alerts if a particular perimeter is breached from where their loved one is supposed to be staying [8].

### 4.4 Use Case: Retail Shopping

Worth \$2.6 trillion, the United States retail industry comprises 15% of national gross domestic product [13]. The ground is shifting underneath this industry though as brick and mortar stores are under siege from a surging market share by Amazon, which is up 150% since 2013.

These traditional stores still hold the top rankings in the retail sales by size, but the ability of Amazon to utilize Big Data for a personalized shopping experience online is forcing these top retailers to adapt with a competing level of customization. Amazon's recommendation engine allows them to see into a user's purchase history, viewing history, rating history and search history, which are all used to point the customer to the most likely product they're looking for. In fact, Amazon is even working on an IoT sensor that they intend will act as a personalized stylist. The device will take a picture of your outfit and make recommendations of what would look best, based on the recommendations of its algorithms that are supplemented by fashion stylists to reflect current trends [33]. As a result, IoT gives Amazon a level of scalability to its entire customer base to create more information and data about the customer that is simply not available to the brick and mortar stores.

To try and compete with this personalization though, brick and mortar retailers are using edge computing to introduce technology that was science fiction 15 years ago in the movie Minority Report. In the movie, which takes place in 2054, the main character is rushing through a busy shopping center when he passes various kiosks that address him by name and ask about his recent purchases in the store. This is the reality that retailers are now using through real-time facial recognition, enabled by edge computing to integrate IoT and Big Data. With this, they are also collecting broader demographic statistics by tracking customers' ages, ethnicity and gender [32]. In fact, America's largest retailer, Wal-Mart, is currently using facial technology to sense customer's moods and find those who are dissatisfied [40].

While we haven't quite hit the personalization depicted in Minority Report for the general public, those with celebrity can expect that high-end stores they visit will recognize them upon entry. For example, one such jewelry store in Los Angeles is equipped with facial recognition technology, stocked with a database of celebrity pictures from Google Images and when someone is recognized, an alert is sent to the manager with purchase history and sizes [46].

Outside of custom shopping, facial recognition is also being used by traditional stores to deal with a risk that e-commerce is not exposed to - shoplifting. With this technology, a retail store can identify when a known shoplifter is most likely to re-visit the store and when, which were previously unquantifiable. Once they are identified on site, management is sent an instant alert and the customer is escorted from the store to prevent further loss in the future [11]. Additionally, RFID sensors are being used on items individually to better track items outside of the store for loss prevention like this and better supply chain management [10].

## 5 CONCERN WITH IOT

As exciting as these use cases are about what the future might hold, innovation is outpacing legislation for IoT. As we will expand upon below, a race to release products has left consumers susceptible to hacking in some cases as security measures have not been fully developed yet for these devices. Additionally, with the customization that comes with IoT, consumer information is being sold to advertising agencies in many cases without the consumer's knowledge.

## 5.1 Security

While we have discussed some of the most exciting and interesting developments in IoT, this blistering pace of innovation has come at a price. There are experts in this field that believe the connectivity of these devices are a gateway of vulnerability as many IoT devices do not have sufficient security measures, allowing malicious actors direct access into some of people's most private details.

For most utilizing IoT, the technology is used to make their lives easier in some respect. However, when it comes to security, it is believed this approach of a "hands off" relationship with IoT leaves users susceptible to security breaches. Specifically, users need to be diligent in making sure their software is up to date across *all* devices. The reason for this is that with a large network of IoT devices, hackers now have multiple fronts on which they get behind the firewall whereas their only avenues traditionally were the computer and more recently, smart phones. As a result, negligence in one area could be enough of an opening for a comprised network where hackers could take control of a device, which is particularly worrisome in the case of an autonomous car.

Another dimension of risk for IoT security sits with the creators of this technology. Underlying in the assumption about users being diligent in updating their software to prevent breaches is that the developers of the software are actually making continuous updates to adapt along with hackers. However, as time goes on, new products are likely to draw a company's limited resources away from maintaining older products.

In response to these risks, two significant changes have been undertaken to mitigate some of the risks. Namely, companies have introduced automatic updates and used the same operating system across later models of a particular product. These automatic updates then take the burden off of the user of IoT technology, which is an attractive feature as many adopt the technology to simplify their daily life. Additionally, when companies are able to use the same underlying operating system across later products, they're able to update all products in lockstep with the developing security community, ensuring no older products are left behind as an opening behind the firewall [39].

Fortunately, these security concerns with IoT have largely played out in the hypothetical. In fact, surveys have found that the majority of consumers are unaware of IoT security risks and once made aware, do not consider the risks serious. In fact, surveyors even found that if a device had a known security flaw, 20% of consumers are still willing to buy the product [9].

For this reason, with no major attacks to date, adopters of IoT have possibly felt insulated as an overwhelming majority are not threatened by the security risks IoT could pose. This is not to insinuate that IoT attacks do not regularly happen, but instead that they have not occurred on the scale that some of the largest security breaches in recent years have occurred, such as the Target Corporation's incident in 2013. In that breach, 110 million consumer credit card numbers were stolen, along with personally identifying information like their address, e-mail and phone number. The entire episode was estimated to have cost Target \$162 million [2].

While an IoT originated attack like this has not happened yet on this scale, these attacks do occur with frequency. One such statistic demonstrating this unsettling fact is that half of all companies that

have adopted some element of IoT technology have experienced a security breach. In the end, these breaches have cost an average of 13% of annual revenue [45].

The closest demonstration of IoT risks came in October 2016 through the "Mirai" malware, which was used to attack DNS servers and bring down high traffic websites, such as Netflix and Amazon. Disturbingly, "Mirai" translates to "future" in Japanese. With Mirai, the program is continuously scanning the internet for IoT connected devices that have left the default user name and password. Then, once a device is found, it is turned into a bot that is used to amplify a DDoS attack. Incredibly, the average IoT device is scanned every two minutes with this bot, leaving an extremely small margin for error in being compromised [44].

This breach demonstrated the downside of the highly connected nature of IoT. Against the benefit of having devices that can communicate with one another, in the event of an attack, these devices are intertwined and will be equally compromised. The network of IoT devices has gotten so complicated for some companies that one survey found 66% of IT professionals aren't sure how many devices are in their environment [35].

## 5.2 Privacy

It is rather commonplace knowledge, for better or worse, that the apps we use daily are collecting data on us. We're aware that it is on going, but in many cases it's unclear what data is actually being collected. This data aggregation is one of the main debates around IoT. Ironically, one of IoT's primary benefits makes it also one of the most unsettling for others, fearing how the data could be used in the wrong hands. In fact, in 2014, it was found that of the top 200 free apps in the Apple store, 95% were engaging in "risky behavior" [36]. These risky behaviors, are defined as activities such as tracking locations, accessing users' contact lists or selling registration data to ad agencies.

Due to this pervasive data collection, one of the consequences of a security breach via an IoT device would be having personal information compromised. However, outside of this direct relationship, there are concerns on privacy as it relates to usage as laws are behind technology in how this data can be used. The only major pieces of legislation that concern privacy at the federal level are through HIPAA for medical records and the Fair Credit and Reporting Act. Outside of these, the task of regulating privacy is left to states, which are behind the curve in today's fast paced, data driven world.

In a similar conundrum as the security concerns with IoT, one of its greatest features in its ability to continuously monitor and collect this data into Big Data sets is also the reason some hold reservations on the technology. This is mainly due to the fact that this data is not collected into a central repository, like your credit, to see what information is being associated with you. To take it a step further, it is not even clear who has what data on a particular user.

In a shock to most on how little personal privacy may exist in our technology saturated world, it was discovered that the CIA and MI-5 intelligence agencies were using "smart" TVs to eavesdrop on conversations in people's homes. For security experts, this was no surprise and known to be an easily accessible device, but

those outside of that community felt an invasion of privacy [49]. Discovered in 2016, the program was used in 2014 by exploiting the voice enabled features that Samsung included in its TVs to listen to conversations. The power button was even programmed to look as if the TV was off while this recording was happening [56].

While this spying was alleged to have just been on “people of interest”, the average consumer with a smart TV has likely experienced spying they were unaware of through their viewing habits. By default, Vizio TVs were found to be recording their customers activities by logging metrics such as date, time, show, whether it was live or recorded and how long it was watched. This is estimated to have affected 11 million TVs in the end before the FTC outlawed the practice of having these settings turned on by default [27]. This would be a utilization of IoT and Big Data that few would be comfortable forfeiting without their consent.

## 6 CONCLUSION

As we've seen, IoT cannot realize its full potential without Big Data. The IoT universe represents the senses by which Big Data is collected for later insights and innovations. For this reason, the IoT revolution has the potential to completely change the world as we currently know. It could be a world in which automobile accidents are no longer a tragic reality or a world where health care delivers the most personalized plan with attention on every minute detail. Additionally, users are able to benefit from the increase in computing power per dollar spent, allowing them more flexibility than ever to design their own IoT device, as was demonstrated in the application made for this paper. However, against this rapid pace of innovation in IoT, some of its most attractive features of interdependency among devices expose the technology to some of its greatest vulnerabilities. Keeping this growth rate in the products in step with security will prove to be one of the biggest challenges in coming years.

## A CODE COMPILATION AND SAMPLE OUTPUT

The following urls are intended to direct to various parts of the project.

- Packages required to compile the project along with sample input
  - <https://github.com/bigdata-i523/hid334/tree/master/project>
- Python code to create the monitor:
  - <https://github.com/bigdata-i523/blob/master/project/code/project.py>
- Weather codes:
  - <https://github.com/bigdata-i523/blob/master/project/weathercodes.py>
- Kivy file:
  - <https://github.com/bigdata-i523/blob/master/project/DailyView.kv>

## ACKNOWLEDGMENTS

The author would like to thank Professor Dr. Gregor von Laszewski, Juliette Zerick and the other Associate Instructors for their support and suggestions in exploring this topic.

## REFERENCES

- [1] Philip Adams. 2017. Why self-driving cars can't start without edge computing. Website. (07 2017). <https://knect365.com/cloud-enterprise-tech/article/b4751c4b-7b5d-4407-8789-420289799988/autonomous-cars-can't-start-without-edge-computing>
- [2] Taylor Armerding. 2017. The 16 biggest data breaches of the 21st century. (10 2017). <https://www.cscoonline.com/article/2130877/data-breach/the-16-biggest-data-breaches-of-the-21st-century.html>
- [3] Jeremy Barr. 2016. The New York Times Pulls Back Ahead of the Washington Post for Unique Visitors. Website. (02 2016). <http://adage.com/article/media/york-times-pulls-back-ahead-washington-post/302720/>
- [4] Krishna Bharat. 2012. Google News turns 10. Website. (09 2012). <https://blog.google/topics/journalism-news/google-news-turns-10/>
- [5] Johana Bhuiyan. 2017. GMfis self-driving division has hired a former top Uber engineer as its first CTO. Website. (11 2017). <https://www.recode.net/2017/11/30/16720994/gm-cruise-cto-susan-fowler>
- [6] Isaac Christiansen. 2017. The Internet of Things and the Evolution of Elderly Care. Website. (06 2017). <http://www.iotevolutionworld.com/smart-home/articles/432936-internet-things-the-evolution-elderly-care.htm>
- [7] Michael Coren. 2016. Tesla has 780 million miles of driving data, and adds another million every 10 hours. Website. (05 2016). <https://qz.com/694520/tesla-has-780-million-miles-of-driving-data-and-adds-another-million-every-10-hours/>
- [8] Reenita Das. 2017. 10 Ways The Internet of Medical Things Is Revolutionizing Senior Care. (05 2017). <https://www.forbes.com/sites/reenitadas/2017/05/22/10-ways-internet-of-medical-things-is-revolutionizing-senior-care/#5e01a7965cf>
- [9] Gary Davis. 2017. A Cybersecurity Carol: Key Takeaways From This Year's Most Hackable Holiday Gifts. Website. (11 2017). <https://securingtomorrow.mcafee.com/consumer/consumer-threat-notices/most-hackable-gifts/>
- [10] Jim Donaldson. 2016. Why Retailers Are Turning To RFID For Loss Prevention. Website. (Aug. 2016). <https://www.mojix.com/retailers-rfid-loss-prevention/>
- [11] The Daily Dose. 2017. Stopping Shoplifters Goes High-Tech. Website. (June 2017). <http://www.ozy.com/fast-forward/stopping-shoplifters-goes-high-tech/78920>
- [12] Robert Epstein. 2016. The New Censorship. Website. (06 2016). <https://www.usnews.com/opinion/articles/2016-06-22/google-is-the-worlds-biggest-censor-and-its-power-must-be-regulated>
- [13] National Retail Federation. 2017. The Economic Impact of the U.S. Retail Industry. Website. (2017). <https://nrf.com/resources/retail-library/the-economic-impact-of-the-us-retail-industry>
- [14] FICO. 2017. Behavioral Analytics Attack Fraud, Cyber and Financial Crime. (04 2017). <http://www.fico.com/en/blogs/analytics-optimization/behavioral-analytics-for-fraud-cyber-and-financial-crime/>
- [15] Frederic Filloux. 2013. Google News: the secret sauce. Website. (02 2013). <https://www.theguardian.com/technology/2013/feb/25/1>
- [16] Arik Gabbai. 2015. Kevin Ashton Describes fithe Internet of Thingsf. Magazine. (01 2015). <https://www.smithsonianmag.com/innovation/kevin-ashton-describes-the-internet-of-things-180953749/>
- [17] Garmin. 2017. Garmin Edge 500. Website. (2017). <https://buy.garmin.com/en-US/US/p/36728#overview>
- [18] Gartner. 2017. Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016. (02 2017).
- [19] Christian Gilbertson. 2017. Here's How The Sensors in Autonomous Cars Work. Website. (03 2017). <http://www.thedrive.com/tech/8657/heres-how-the-sensors-in-autonomous-cars-work>
- [20] Google. 2017. Mobile App Reporting in Google Analytics - iOS. Website. (2017). [https://developers.google.com/analytics/devguides/collection/firebase/ios/#how\\_does\\_it\\_work](https://developers.google.com/analytics/devguides/collection/firebase/ios/#how_does_it_work)
- [21] Andrew Hawkins. 2017. Nvidia says its new supercomputer will enable the highest level of automated driving. Website. (10 2017). <https://www.theverge.com/2017/10/10/16449416/nvidia-pegasus-self-driving-car-ai-robotaxi>
- [22] Patrick Hoge. 2010. Survey: 44% stop at Google News headlines. Website. (01 2010). <https://www.bizjournals.com/sanfrancisco/stories/2010/01/18/daily24.html>
- [23] Nabeel Hyatt. 2017. Autonomous driving is here, and it's going to change everything. Website. (04 2017). <https://www.recode.net/2017/4/19/15364608/autonomous-self-driving-cars-impact-disruption-society-mobility>
- [24] IBM. 2015. IBM Plans to Acquire The Weather Company's Product and Technology Businesses; Extends Power of Watson to the Internet of Things. Press Release. (10 2015). <http://www-03.ibm.com/press/us/en/pressrelease/47952.wss>
- [25] Ajit Jaokar. 2017. Behavioural Biometrics, IoT and AI. Website. (10 2017). <https://www.datasciencecentral.com/profiles/blogs/behavioural-biometrics-iot-and-ai>
- [26] Search Engine Journal. 2016. Over 60% of People Trust Google for News vs. Actual News Sources. Website. (01 2016). <https://www.searchenginejournal.com/google-news-2/154475/>
- [27] Jacob Kastrenakes. 2017. Most smart TVs are tracking you fi? Vizio just got caught. (02 2017). <https://www.theverge.com/2017/2/7/14527360/>

- vizio-smart-tv-tracking-settlement-disable-settings
- [28] Suzanne Kattau. 2015. Research from Gartner: Real-Time Analytics with the Internet of Things. Website. (06 2015). <https://www.rtinsights.com/research-from-gartner-real-time-analytics-with-the-internet-of-things-dw/>
- [29] John Kiernan. 2017. Credit Card & Debit Card Fraud Statistics. Website. (02 2017). <https://wallethub.com/edu/credit-debit-card-fraud-statistics/25725/>
- [30] Sam Levin and Mark Harris. 2017. The road ahead: self-driving cars on the brink of a revolution in California. Website. (03 2017). <https://www.theguardian.com/technology/2017/mar/17/self-driving-cars-california-regulation-google-uber-tesla>
- [31] Karen Lewis. 2017. Visa and IBM are bringing the world secure payment experiences through the IoT. (02 2017). <https://www.ibm.com/blogs/internet-of-things/visa/>
- [32] Annie Lin. 2017. Facial recognition is tracking customers as they shop in stores, tech company says. Website. (11 2017). <https://www.cnbc.com/2017/11/23/facial-recognition-is-tracking-customers-as-they-shop-in-stores-tech-company-says.html>
- [33] Jon Markman. 2017. Amazon Using AI, Big Data To Accelerate Profits. Website. (06 2017). <https://www.forbes.com/sites/jonmarkman/2017/06/05/amazon-using-ai-big-data-to-accelerate-profits/#12f29cb6d55>
- [34] Mark Mather. 2016. Fact Sheet: Aging in the United States. Media Guide. (01 2016). <http://www.prb.org/Publications/Media-Guides/2016/aging-unitedstates-fact-sheet.aspx>
- [35] Kayla Matthews. 2017. 4 Statistics That Reveal Major Problems With IoT Security. Website. (02 2017). <https://channels.theinnovationenterprise.com/articles/4-statistics-that-reveal-major-problems-with-iot-security>
- [36] Neil McAllister. 2014. How many mobile apps collect data on users? Oh ... nearly all of them. Website. (02 2014). [https://www.theregister.co.uk/2014/02/21/appthority\\_app-privacy\\_study/](https://www.theregister.co.uk/2014/02/21/appthority_app-privacy_study/)
- [37] Microsoft. 2017. Five ways edge computing will transform business. Website. (09 2017). <https://blogs.microsoft.com/iot/2017/09/19/five-ways-edge-computing-will-transform-business/>
- [38] Patrick Nelson. 2016. Just one autonomous car will use 4,000 GB of data/day. Website. (12 2016). <https://www.networkworld.com/article/3147892/internet/one-autonomous-car-will-use-4000-gb-of-datataday.html>
- [39] University of Missouri System. 2016. Securing the Internet of Things (IoT). Website. (11 2016). [https://www.umsystem.edu/makeitsafe/securing\\_the\\_internet\\_of\\_things\\_iot](https://www.umsystem.edu/makeitsafe/securing_the_internet_of_things_iot)
- [40] Dan O'Shea. 2017. Report: Walmart developing facial-recognition tech. Website. (07 2017). <https://www.retaildive.com/news/report-walmart-developing-facial-recognition-tech/447478/>
- [41] Kasey Panetta. 2017. Gartner Top Strategic Predictions for 2018 and Beyond. Website. (10 2017). <https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/>
- [42] Nevon Projects. 2017. IOT Heart Attack Detection & Heart Rate Monitor. Website. (2017). <http://nevonprojects.com/iot-heart-attack-detection-heart-rate-monitor/>
- [43] Lopez Research. 2013. An Introduction to the Internet of Things (IoT). Research Report. (11 2013). [https://www.cisco.com/c/dam/en\\_us/solutions/trends/iot/introduction\\_to\\_IoT\\_november.pdf](https://www.cisco.com/c/dam/en_us/solutions/trends/iot/introduction_to_IoT_november.pdf)
- [44] Symantec Security Response. 2016. Mirai: what you need to know about the botnet behind recent major DDoS attacks. Website. (10 2016). <https://www.symantec.com/connect/blogs/mirai-what-you-need-know-about-botnet-behind-recent-major-ddos-attacks>
- [45] Freddie Roberts. 2017. Half of US companies hit by IoT security breaches, says survey. (06 2017). <https://internetofbusiness.com/half-us-iot-security-breach/>
- [46] Brenda Salinas. 2013. High-End Stores Use Facial Recognition Tools To Spot VIPs. Website. (07 2013).
- [47] Antony Savvas. 2014. The Weather Company turns to open source big data analytics. Website. (11 2014). <https://www.computerworlduk.com/data/kpmg-launches-big-data-investment-fund-3489089/>
- [48] Claire Scholz. 2015. Biometrics to Secure the Internet of Things. Website. (12 2015). <https://blog.bioconnect.com/2552/biometrics-to-secure-the-internet-of-things/>
- [49] Stilgherrian. 2013. Smart TVs are dumb, and so are we. Website. (10 2013). <http://www.zdnet.com/article/smart-tvs-are-dumb-and-so-are-we/>
- [50] Mark Sullivan. 2012. A brief history of GPS. Website. (08 2012). <https://www.pcworld.com/article/2000276/a-brief-history-of-gps.html>
- [51] Raj Talluri. 2017. Why edge computing is critical for the IoT. Website. (10 2017). <https://www.networkworld.com/article/3234708/internet-of-things/why-edge-computing-is-critical-for-the-iot.html>
- [52] Versa Technology. 2017. How much Data will The Internet of Things (IoT) Generate by 2020? Website. (10 2017). <https://www.versatek.com/blog/how-much-data-will-the-internet-of-things-iot-generate-by-2020/>
- [53] Heather Thompson. 2017. Edge computing: It's what healthcare IoT craves. Website. (03 2017). <http://www.medicaldesignandsourcing.com/edge-computing-healthcare-iot-craves/>
- [54] Weather Underground. 2017. Weather Underground - About Our Data. Website. (2017). <https://www.wunderground.com/about/data>

## 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

[2017-12-10 13.52.31] pdflatex report.tex

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

```
name: Peter Russell
hid: 334
paper1: Oct 28 17 100%
paper2: Nov 24 17 100%
project: Dec 04 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
8
wc 334 project 8 6920 report.tex
wc 334 project 8 7546 report.pdf
wc 334 project 8 1887 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

```
4: \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
```

```
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 and Its Application in Education

Weipeng Yang

School of Education, Indiana University Bloomington  
201 N Rose Ave  
Bloomington, Indiana 47405  
yang306@umail.iu.edu

## ABSTRACT

The development of big data is changing the society in a dynamic way. With high speed internet and high penetration rate of mobile devices, every individual becomes a source of data and is constantly provide these data to various organizations who want to make profits or want to utilize these data to contribute to a certain end. Big data helps online retailers to add new strategies to increase their selling; it also helps medicare organizations to make more accurate diagnosis to patients; it even changes the sports industry. Because big data is changes various industries profoundly, it will certainly in a point change the way people acquire new knowledge. For this reason, it is important to review how big data bring changes to different industries and how education strategies could adjust to the fast-change world. In order to have a clear picture of in what ways big data will influence educational strategies, we will use education recommendation system and medicare education which is transformed by big data, as a case to see how educators should adjust their strategies with the benefits brought by big data..

## KEYWORDS

i523, HID236,HID218, big data, education

## 1 INTRODUCTION

When we search the definition of education, we will find that it is defined as process of facilitating learning or the acquisition of knowledge, skills, values and beliefs and habits [43]. However, when we talk about education, we regard education as academic education such as K-12 education and higher education. In reality, education is happening all the time and everywhere. When you are sitting in a classroom, you are learning to get a degree and plan for your career life. When you are working in the office, you gradually learn how to accommodate what you learn at school to what the reality is in the workplace. When you are watching TV, you are getting information about what is happening around the world and you gain your first impression of different countries around the globe. And even when you are shopping online, you are learning how to identify the product is good or bad by reading the reviews. Therefore, in the 21st century when educators look into education, it will cover not only school education but also corporate training and other forms of informal learning.

## 2 A LOOK BACK AT LEARNING

It is very difficult not to think of Confucius and Socrates, two great men who are regarded as the most important people in terms of influence on education in the East and West. Confucius emphasized the importance of education and proposed that education should be equal to everyone. He was a teacher himself and taught students

Geng Niu

School of Education, Indiana University Bloomington  
201 N Rose Ave  
Bloomington, Indiana 47405  
Niugeng@umail.iu.edu

morality, proper speech, government and refined arts. "He never discourses at length on a subject. Instead he poses questions, cites passages from the classics, or uses apt analogies, and waits for his students to arrive at the right answers" [30]. Confucius set an example of teachers in ancient China and other regions in Asia such Korea and Japan. Teachers in these parts of world should be much better learned than students and are examples in terms of morals. Because of this, students should imitate teachers. By contrast, Socrates adopted a different view of learning. fSocrates does not believe that any one person or any one school of thought is authoritative or has the wisdom to teach "things." Socrates repeatedly disavows his own knowledge and his own methods. However, this appears to be a technique for engaging others and empowering the conversator to openly Dialogue [2]. This may be one of the reason why the learning style in the East and the West is so different. But when we look back at education in schools in the past, no matter where the school or the learning place is located, learning is highly teacher-centered. In ancient China, teachers still follow the learning style of confucius. The teachers had authority over their students and students were supposed to treat their teachers like their fathers. In return, teachers should be selfless enough to pass what they knew to their students. In the West, the situation was similar. And one of the reasons might be the lack of resources in teaching. In ancient times, only limited number of books were available due to printing technique and the number of scholars who could write books. In the year of 1500, the illiteracy rate of men and women in English is 90 and 100 respectively. And in Qing Dynasty around 1880, literacy rate is around 30-45 in men and 2-10 in women. It is not hard to deduce that these numbers are much lower in 1500s [28]. Knowledge or information was in the hand of the top 10 of people which make knowledge more precious. Therefore, teaching must be teacher-centered and teachers have much authority over the students. However, it is never true today. The volume of books will take a person's whole life to read; TV has changed the way of how people get information; and the internet revolutionizes how information is created and transmitted. As a result, teachers are no longer just knowledge providers and it is impossible for them to be mere knowledge providers because learners have access to almost infinite source of knowledge.

## 3 WHAT IS INSTRUCTIONAL DESIGN?

"Educational technology is the study and ethical practice of facilitating learning and improving performance by creating, using, and managing appropriate technological processes and resources" [13]. Instructional technology started from instructional media , the use of which can date back to the first 10 years of 20th century in school museums. The use of different media in instruction or

learning have gone through visual instruction, audio-visual instruction and the use of communication theories to today the integration of computers and internet technologies [29].

However, the turning point of the birth of educational technology began as visual education. At the turn of the 20th century, educators were exploring the potentials of motion pictures and projected slides. In the 1950s, the advent of television added new dimension of widespread of audio-visual programming. At this time, the design materials only focused on creating attractive and creative presentations which are pleasing to learners' eye and ears. But a shift happened in the next decade. Educators not only cared about the appeal of the teaching or learning materials, but also cared about what learners are doing. In the next a few decades, the focus of learning design continued to change because of the advent of the internet which allows learners to collaborate anytime anywhere. Also, computers became a powerful assistant in learning with the advances made in CPU and storage [25].

Instructional design has several names such as instructional system technologies, learning design and educational technology. Although universities which have program on instructional design prefer different names, they have similar courses and goals of training. The definition of instructional design has been revised several times in history and those changes were caused by different opinions held by experts in the field and most important caused by advances of science and technology.

#### **4 THE DEVELOPMENT OF EDUCATION PSYCHOLOGY**

Behaviorists believe that performance of people can be changed by contingencies of reinforcement combined with changes in the environment [42]. For example, drivers or passengers of a car may not want to or forget to fasten the safety belts. In order to prevent that from happening, a machine would give out a loud "beep" noise which is annoying to tell people in the car that you need to fasten the belt. This is called a negative reinforcement. In order to avoid something awful to occur, people will behave in certain way. Opposite to this is positive reinforcement. For instance, teachers would give a student who has the high scores in a exam a gift and verbal appraise as a way to encourage all the students to study harder. Behaviorism has great impact on programmed instruction in both academic education and military training [34]. However, the behaviorist approach to learning has two problems. The first one is the use of proper reinforcement. As learners grow older, instructors have to find reinforcement that learners will value. But it is really difficult to provide such reinforcement to adult learners. The other problem is that behaviorists seem to ignore the process of learning.

By contrast, cognitivists focus on how people process information. The core components of cognitive approach to learning are perception and sensory stores, short-term memory, and long-term memory. Perception is about how people select what information to pay attention to; "sensory stores are capable of storing almost complete records of what we attend to but hold those records very briefly"; short-term memory helps people to rehearse the information coming through sensory stores but it has limited capacity; Long-term memory is where information is stored in a certain

way permanently and is ready to be retrieved [41]. A example of using cognitive theory in learning or information is the design of presentation slides. A good PowerPoint presentation may contains clear contrast between the texts and the background or different categories of information. It also only contains keywords of a topic so viewers will be well-guided by the slides when listening to the speakers. Although cognitive approach of learning helps learners to process information, what learners can do if they need to learn a ill-defined topic?

Constructivism made one more step forward towards learning. Learning, according to constructivist theory, is a process of meaning making, a process of solving problems when encountering cognitive conflict and a social activity such as collaboration and negotiation [44] Put it simply, constructive theory advocates simulation of the real world environment. A topic is ill-defined and learners are required to formulate their own strategies to look for relevant information and experiment potential solutions to solve a problem.

#### **5 PROBLEM-BASED LEARNING AND MEDICAL EDUCATION:**

Problem-based learning is based on constructive approach to learning. "Participation in valued activities within different domains is fundamental to how students learn." People who advocate this problem-based approach suggest that learning happens when other people involve such peers, tutors or mentors. And cooperation in activities can lead to higher reasoning level. Students may change their perspective of thinking and their opinions about a topic because in collaboration more ideas are involved and those ideas will be discussed in an environment in which sharing and collaboration are promoted. Lev Vygotsky found the construct of the zone of proximal development to explain how people can facilitate knowledge construction. This framework shows that if instructors can reduce the distance between what the learners can do completely by themselves and the things that can be accomplished by themselves with the assistance from others, then the instruction can be successful. "PBL is a form of education in which information is mastered in the same context in which it will be used" [6]. Another definition of PBL is "a learning method based on the principle of using problems as a starting point for the acquisition and integration of new knowledge" [23]. Problem-based learning can also refers to problem-and -task-centered approaches of learning. It is one of educational technologies designed to situate instruction in authentic or meaningful settings. It has been employed in many different fields of studies such as medicine, science, law, business and mathematics. And the goals of PBL differs from each other. In medicine PBL requires learners to work in groups to practice their skills to diagnose patient cases and the ability to use clinical knowledge in practice. But in science and humanities, students in a PBL class to come up with explanations for a certain phenomenon through activities such as defining a question, seeking evidence, and outlining and argument. Moreover, in law and business related courses students will engage in the study of cases and they will be encouraged to seek and summarize critical information from those cases and present their finds to peers in the classroom. By the end of their presentation, instructors will provide feedback. In math and science PBL courses, students work together in an environment in

which constructive feedback is provided to each other or by tutors or teachers. Although the goals of learning in different fields are different, in PBL the core value is to put education in authentic tasks so the learning is more meaningful.

Savery and Duffy proposed a framework of how to conduct problem-based learning: Anchor all learning activities to a larger task or problem

Support the learner in developing ownership for the overall problem or task

Design an authentic task

Design the task and the learning environment to reflect the complexity of the environment they should be able to function in at the end of learning

Give the learner ownership of the process used to develop a solution

Design the learning environment to support and challenge the learner's thinking

Encourage testing ideas against alternative views and alternative contexts

Provide opportunity for and support reflection on both the content learned and the learning process [35]

Medical education is very suitable for Problem-based learning because the advances made in medicare makes it impossible to include everything in lectures. And in the field of medicare, doctors will face various problems with patients which are highly likely beyond what medical students can learn from school. Therefore, in order to foster the ability to solve problems, critical thinking and experiment potential solutions, PBL serves as a critical part of medical education. PBL has been employed in many medical schools around the world. "It was introduced in the medical school at Mc-Master University in Canada in the late 1960s and is now a common curriculum component in medical and health science schools around the world" [23]. "The University of New Mexico was the first to adopt a medical PBL curriculum in the United States and Mercer University School of Medicine in Georgia was the first U.S medical school to employ PBL as its only curricular offering" [6].

Here we will present how to use this framework with medical education. The first step of designing PBL for medical students is to find an authentic task. By saying authentic task, we don't mean the task must be the same with what happens in a hospital every day. It means the task will require similar cognitive load to a real problem. The specific difficulty of the task is designed by the instructor according to the level of the course. As is often the case, instructors will create scenarios to represent an authentic task. Before creating a scenario, instructors should formulate objectives of the course, and create a scenario in which all of these objectives will be accomplished. The complexity or the difficulty of the problem should be appropriate to the curriculum and the level of students' understanding. It is better if the scenario is appealing enough to attract students' attention. Basic science should be included in the context of a clinical scenario to encourage integration of knowledge. Although the problem presented in a PBL class should be ill-defined, the PBL scenario should have cues to stimulate discussion and push students to seek reasonable explanation to the issues involved in the scenario. At last, the scenarios should promote participation by

the students in the seek of explanation [46].

The next step is to gather all relevant information which can be useful or useless to the final solution to the diagnosis. However, this information will not be provided to students directly. Besides relevant information, the instructor should also have other resources such as equipments, lab or simulation of the environment in a real hospital. Then students will have a meeting with the instructor to talk about the basic information of the patient such as age and gender and symptoms he or she has. Then the task will almost completely hand to the students. Here comes to an important point of the whole learning experience. The learners should be told that there is no correct answer to the task and it is the students' responsibility to find a possible solutions.

A tutor should be assigned to the students and the tutor is not necessarily an expert in medicare because the tutor's job is not to provide suggestions to the students. Or the instructor can be the tutor. But the responsibility of the tutor is to ask leading questions such as why do you choose this? or how did this happen? By asking these questions we hope students will spot their own mistakes or loopholes in the process of finding a solution to the diagnosis. In this way students will revise their strategy of work. Another responsibility of the tutor or instructor is to provide key information if the students are seriously off the track. And in order to ensure the quality of the course, tutors or instructors have to do so. Speaking of the roles of tutor in PBL, we have to talk about scaffolding. A real or physical scaffold is a structure to support learners to complete a task and it is not permanent. When a task is accomplished, the structure will be removed. It is still the same when we talk about scaffolding in education. It will be removed when it is not needed. Scaffolding is designed to assist learners to complete tasks which are otherwise beyond their reach. This suggests that the design of scaffolding must be very careful. So there are several questions for tutor and instructors to think when they design such a structure: what is needed to support, when and in what way to support the students, how much support should be provided to learners, and when and how to fade scaffolding.

In PBL in a medicare course, students are required to write reports weekly or bi-weekly on how they collaborate, what problems occur and how they solve these problems. Also, by knowing the progress students make, it is much easier for instructors to see how they grow and how they should adjust some elements of the learning environment. And a final report to summarize the whole process of collaboration and the working process will be submitted at the end of the semester.

Such courses can also be conducted in multiple groups. Every group will have their own way of collaboration and propose different solutions to a diagnosis. And instructors should create an environment where different groups are eager to share their own progress because they can always get constructive feedback from their peers. Moreover, such a sharing environment will make the learning more dynamic and accelerate the growth of learners. Tutorials are also an important element in PBL. Usually, the PBL tutorial has a group of students which has no more than 10 and a tutor who provides scaffolding to the session. The duration of the session varies. It depends on how long it takes for a certain group to have good dynamics. Moreover, for each tutorial session, a different

leader or chair should be elected so every member of the team will contribute and free ride can be avoided [46].

Here we want to elaborate tools and activities can be included in PBL in the Internet era. Basically, PBL courses can include activities such as generating lists, scaling down the scope of topics, making outlines of options, debating issues, and even voting. Today, many activities can happen in virtual environment. Wikis enable learners to have meeting in a virtual community and collaborate on projects and solve problems. And meeting tools such as Zoom, Goggle Hangouts and Adobe Connect enable online meetings of a large group of students and share screens and notes. Moreover, blogs also provide virtual space for learners to practice their writing skills and share their writing with audiences beyond their teacher. In a PBL class, web 2.0 tools can also be included such as Skype, Twitter, Instagram.

Here are some examples of how Mercer University conducted PBL in its medical courses. At Mercer University, a series of tutorial sessions were used to substitute the lectures. And during each session, faculty members and students would meet to discuss the actual case problems. In other programs which are related to clinical skills and community science, students need to deal with simulated patients and spend some afternoons with local primary-care practitioners. "In this way, real life clinical practice in a rural community becomes a laboratory exercise for the illustration of basic science theory." In Mercer University, tutors were called "faculty overseers" who are neither to be the source of all information nor even to have information about every area being discussed. The responsibility of these overseers is to keep student participation and knows enough to prevent gross mistakes. On the contrary, students were teachers and learners. Without giving lists of what to know, students need to generate a list of what to look for according to importance of relevant information.

Although small groups of meeting played essential role in the PBL of Mercer University, lectures are still used to some extent. The students may have some lectures on one or two basic science lectures every week but these sessions were not mandatory. The evaluation of the course was intense. Students at Mercer University were tested by both intramural and extramural means. At the end of each of the thirteen curricular phases, students would have a 200-item, cross-disciplinary, objective examination and a forty minutes case analysis oral examination.

The majority of faculty members favor PBL over the conventional way of teaching. The reason is very simple: it is a more natural way of learning. PBL simulate the environment where people generate knowledge. For example, students became better prepared in the learning process. That is the ownership was handed to the students instead of the instructors. If a student came to the discussion session without any preparation, she or he would be complaint by other members. Another benefit of PBL is that students became more flexible in learning. Students at Mercer University used texts, mono graphs, periodical literature and various resources in their learning. In the past, the learning is very lecture centered and students were actually not actively engaged in the learning. By contrast, when they were on their own, they tried every alternatives to find useful

resources and developed flexibility in learning [23].

From the history of the evolution of educational technology we can see the changes are brought by technological development made in other fields of studies. Those technologies were not intended to contribute to education but they are all utilized in education. And to successfully employ PBL in an academic learning environment, instructors and instructional designers must build a proper environment. As a result, the development of big data can provide new thoughts in how to advance current instructional design and improve the building of a proper PBL learning environment.

## 6 CHALLENGES OF LEARNING IN THE INFORMATION ERA

The challenges of learning in 21 first century is that the explosion of information brought too much information whose credibility is uncertain. Many people, especially scholars, questions the accuracy of information of Wikipedia. However, Wikipedia may be the most popular sites for all kinds of information ranging from entertainment to academia. And because of the affordable and high speed internet, everyone has a say in the virtual world. One can find people argue on an issue in online forums, express their own opinions in blogs and social media. However, these information could be wrong and there is no third party to verify if the information is correct. By the end, people tend to believe in the opinions presented by the most popular sites or people. For example, in China a high school history teacher go visual on the internet and he starts to have his own online courses about history in China and other parts of the world. His courses are pretty interesting because a lot of humor is involved and various media are used such as animation and movies. Therefore, a lot of students prefer to watch his online courses instead of taking the face-to-face class at school. However, the opinions presented by this history teacher are very different from main-stream scholars especially in the history of the second world war and civil war in China. And this caused problems at schools. In this case I presented, the teacher actually unconsciously took advantage of the populism of teenagers at high school. Students at this age can be very disobedient and do not want to engage in the old tradition.

And here comes another problem and the internet era. It is often the case that who has the most resources to populate a opinion will finally be the person who has the most say. It seems that the internet give people equal opportunity to express. However, what really happens is that people can only find limited opinions or values. For examples, many news agency can use the resources they have to control media on what to be reported and what not to be reported. The fake news of several US news agency proves that it is real. In addition, the internet world is actually not so different from the physical reality. One is likely to find that the best resources on the internet are also expensive and only open to a few instead of the public general. That is also one of the reason why Wikipedia can be so popular because it is free to everyone. Because the best resources are only open to a small group of people that may widening the gap between the well-educated and the ill-educated. That is also the reason why the education community are working on Open

**Education Resources.** Work with people from the academia, these open resources can be affordable or even completely free and still they have high quality. Massive Open Online Courses can be viewed as the most popular representatives of OER. However, there is still a long way to go in promoting education equality due to political and financial reasons.

**Challenges to instructional design** The last challenge is how to do a thorough analysis of learning. In instructional design, ADDIE model is the most used model of doing the design process. ADDIE stands for analysis, design, development and evaluation. In the analysis phase, instructional designers need to work with subject-matter experts to formulate learning objectives. The learning objectives are specific performance which can be observed or evaluated in other ways. And learner analysis will include the traits of learners, the learning styles of learners, the motivation, confidence, prior knowledge of learners and the potential satisfaction of learners. Also a context analysis will also include. In the design phase, instructional designers will script and finalized learning strategies and tactics for the entire learning experience based on the analysis made in the first phase and the learning materials given by instructors. Then they enter the development phase in which the final education product is made. In the evaluation phase, instructional designers will conduct trials of the course and general a report on what needs to be modified and summative evaluation will be formulated to test learners' performance change in the end.

However, in the internet internet era, the number of online learners can be bigger than 2,000. In many popular MOOCs, there are more than 2,000 people registered. As a consequence, it is impossible to do a learner analysis. Not only the number of students is big, the learning styles, motivations and level of prior knowledge vary drastically. Even if a comprehensive learner analysis is possible, the result might be that the learning environment is too complex and the course may be out of control of the instructors' hand. And in reality it is true. In many MOOC courses, learners have different expectations toward the same course, once they feel disappointed about the course, they drop. And the result is only a very small percentage of learners finally complete the course. And because of the huge number of students, the discussion forum goes out of control and instructors and teaching assistants cannot monitor the discussion and the discussion result in nothing.

## 7 HOW BIG DATA INFLUENCE DIFFERENT INDUSTRIES

Before we look at how big data will influence education or more specifically influence instructional design of medical courses in a PBL environment, we will first examine how big data have influenced other industries. The experience from these industries will provide guidance on how education community utilize big data. Big data has become the buzz words for today's world. One of the reasons is that big data increase benefits of many business. The traditional way of costumer consumption has lasted for centuries. In the ancient time, people would go to fairs to buy groceries, hardware and clothes. But at that time, fairs were not standardized, and the conditions of those fair can be terrible. It was impossible to guarantee the quality of the goods bought by customers. Later, in

the industrialized world, cities were built and shopping mall appeared. In a shopping mall, customers could buy good qualities in different stores. Instead, they would go to supermarket to buy groceries. This mode of doing business remained until the beginning of e-commerce. In the web 2.0 era, search engines enabled consumers to look for products in virtual shops and sellers can collect feedback of consumers' satisfaction in their website [3]. Today, with big data technology, it is possible for online retailors to monitor activities of consumers online. Business owners can have better understanding of consumers and formulate more targeted strategies of how to increase profits. Because of the ability of monitoring online activities and better understanding behaviors of consumers, online retailers can provide personalized services. This is realized through the use of recommender system. Online buyers will be labelled according to their online activities and they will receive emails or suggestions of what to buy on the internet. 35 percent of Amazon's revenue is created by the recommendation system. Users of Amazon can click the recommendation section and see the products selected by the recommendation system. For example, if a learner is looking for a backpack, he or she will probably see some recommended backpack [17]. Dynamic pricing is also a strategy brought by big data technology. "Some business set different prices for their products or services based on algorithms that take into account competitor pricing, supply and demand and other external factors in the market. It is a common practice in industries such as hospitality, travel, entertainment, retail, electricity and public transport" [43]

Before big data was brought to the face of the healthcare system, the role of data in the healing process of patients was minimal. Data such as name, age, disease description, diabetic profile, medical reports and family history of illness were collected. These data could only reflect limited view of a patient. For example, a doctor may know that the reason of a patient with heart disease can be traced back to his or her family, but there are many possible perspectives on why the patient has such disease.(Pal2016) "The influence of big data on medicine is that we can build better health profiles and predictive models around individual patients so that we can better diagnose and treat disease." The pharmaceutical industry is facing the limitation of insufficient understanding of the biology of disease. But big data can help in building the understand of what constitutes a disease such as causes from DNA, proteins and metabolites to cells, tissues, organs, organisms, and ecosystems [36]. The problem for the medical research is that enterprise is unable to follow the pace of the information needs of patients, clinicians, administrators and policy makers. *fiThe flow of new knowledge is too slow, and its scope is too narrow.* The consequence of the medical research community not adopting big data technology is that hospitals are ill prepared for a more precise diagnosis. Now the medical research community need new thinking in their work. The new thinking must involve the integration of new technologies. "For instance, researchers can use big data to reveal clusters of patient groups that might suggest new taxonomies of disease based on how similar they are according to a broad range of characteristics, including outcomes." Advances in prediction can simply attribute to the learning of data and creating a mechanism which is highly reproducible and has consistent performance [18]. "Big data has helped healthcare

institutions take a 360 degree view of a patient's health problems." With the help of big data, new findings, innovative methods of treatment plans and more precise diagnosis can be realized. Here is an example of how it is possible to build better health profiles. Some diseases are more common among a certain race of people due to genetical reasons. When a patient from this race is found suffering from heart disease, the doctors can look at the data of patients belonging to the same race who have same problems. By examining their life style, genetic structure, family DNA and other elements, they can build health profiles for these group of people. Wearable devices can also play a role in the detection of potential health problems even if no apparent symptoms are presented. Wearable devices can help see some indicators of health. And doctors can make certain conclusions and decide on the future action on them. The devices today are already able to record data such as heart rate, pulse, glucose levels and calorie levels. And big data will also have the potential to personalize medicine. The NCI-MATCH trial is examine 1000 people who have tumors that do not respond to standard cancer treatments. Researchers hope that they can match drugs to this kind of tumor to produce the best result [27]. "In the very near future, you could also be sharing this data with your doctor who will use it as part of his or her diagnostic toolbox when you visit them with an ailment. Even if there's nothing wrong with you, access to huge, ever growing databases of information about the state of the health of the general public will allow problems to be spotted before they occur, and remedies - either medicinal or educational - to be prepared in advance" [24].

## 8 HOW BIG DATA WILL INFLUENCE EDUCATION IN GENERAL

The first change we will see in education is the rise of adaptive learning. Adaptive learning means that students can learning knowledge whose difficulty is suitable for their ability. This is enabled by the availability of online application, classroom activity software, social media, blogs and surveys of staff. With adaptive learning comes the universities' ability to provide personalized feedback to students, monitor student satisfaction, increase attainment and give students' opportunities to reflect on their own learning. On the other hand, instructors will receive real-time reports which will enable them to adjust teaching strategies for the best outcomes [20]. Because learning is more adaptive, students can advance their learning in different paces. Big data and data analysts will inform instructors who is learning faster and can advance to a more difficult class and who need support from teachers [14].

Since learning of different learners will at different paces, it is important for learners to develop self-management. For example, in a PBL class, students need to solve an ill-defined problem and the process of learning is almost unguided. As a result, students need to take the initiatives and actively contribute to the project. Also learners will monitor their own process of learning and submit a report to summarize this process. So they must develop their meta-cognitive skills which means the learners are able to learn how to learn. Another reason why self-management is more important in the big data era is the widespread of informal learning. As mentioned before, people today are learning anytime anywhere. Social media, blogs, news and anything connected to the internet

will serve as a source of learning. Therefore, it is impossible for teachers to monitor learning of students all the time.

## 9 BIG DATA MINING

Big data mining refers to the procedure in which a gigantic amount of data from a wide variety of source is collected, and analyzed with a wide spectrum of means to discover inner mechanism or other information via pattern [45]. Being used in almost every field such as business marketing, science and engineering, medicine, design and education industries to provide such functions as intelligence, research and marketing. Oftentimes, big data mining will be carried out on individual persons. When someone is doing activities online, their data will be collected. They could also be providing these data via questionnaires, surveys or other means. This massive amount of data collected on everyone are commonly called big data by the industry and corporations and companies will utilize them to figure out what need one have or what kind of personal trait one may carry. As the big data industry found itself in rapid development, concerns and other critiques are also rising on the ethical issue of big data. Heated discussions were talking about the insult to privacy and abuse of such data. However, big data have already set foot in so many industries and almost all aspects of our daily lives[40].

Data mining sees Artificial Intelligence and Machine Learning as its inception. During data mining, patterns are discovered, and data scientists could utilize such patterns to carry out more versatile functions. The system could get to understand an individual via the data collected about this one. The Recommendation Engine, or so called the Recommender System, is one application for data mining. The recommendation engine filters information and uses data mining techniques to figure out the specific suggestion to one person for information or other assets that may help them with current or future needs.

One commonly used example of the recommender system would be the online shopping websites. When someone shops on it, he or she will be given information about merchandise that related to this purchase. Such recommendations require various kinds of variables, such as this person's shopping history, the gender, age, and occupation of the person, or the items other bought after purchasing the same item. Another example will be after someone searched for a merchandise or service online, the advertisement will pop out for them showing related products.

These kinds of systems require algorithm with high complexity to give out recommendations following patterns discovered via enormous amount of data from mining. Such presentation will be oftentimes beneficial to individuals as they no longer need to go through such amount of information to find their desired service or product. Instead, targeted recommendation will be directly presented to the individual and sometimes the individual will have little awareness that they may need such product or service. As a result, the system will greatly enhance the efficiency for the user, it will also be a blessing for the services or products so that they could be utilized more often.

With such benefits, the recommendation engine is wildly used amongst all online websites, including but not limited to online shopping, searching, streaming and social media websites[39].

## 9.1 Types of Recommendation Engines

A good deal of recommendation engines is backed with such technique called as collaborative or content-based filtering technologies. Collaborative filtering resembles a person making purchases on the gathered information from other via verbal or other means. It could also be understood as crowdsourcing[38]. In many online websites, people could give out ratings or feedbacks for others to reference. It is an interesting phenomenon that customers will more likely to read crowdsourcing comment first rather than the information provided by the seller. The collaborative filtering based system took one step further by categorizing commenters into different subcategories and present different person with different information or resources that might only be beneficial to him or her. Such patterns as statistical models are utilized to calculate everyone's correlation, thus giving out a value of recommendation. Some examples might be Twitter, eBay, Steam and Apple Store. They are all using collaborative filtering systems.

On the other hand, content based systems focus on different properties of a resource, in comparison with the properties of a person. As a person's total using time accumulates, the system will become more and more accurate as the user will demonstrate more personal traits and preferences in using the system. Examples of this kind of content based system will be Netflix and other streaming websites. Moreover, a developed recommendation engine could involve both collaborative and content based filtering techniques to bring prediction accuracy to a new level.

## 9.2 Math Models of the Recommendation Engine

The math models that standing in the back of the data mining engines include such technologies as association, classification and clustering means. Clustering refers to the procedure of combining individual with certain characteristics and trait being recognized as high value in the recommendation system. Such values as ratings, tones of comments are taken weighted average of all members in the cluster to identify the how the individuals in this cluster would recommend this product or service. More complicated systems would involve multiple clusters and calculated overall weighted averages across all the clusters that one individual belongs to[37]. Classification identifying technique are also utilized as the cornerstone of interconnecting different person with different appreciation to different items. Fundamental version of classification systems only works as primary filter to figure out how relevant individuals with desired kind of resources. As an example, only providing infant nutrition food for those who just give birth to a baby. This example only provides a crude vision of classification while more complicated ones will be able to perform prediction recommendations with higher complexity, and thus higher accuracy[33]. Association on the other hand provide more sophisticated recommendation rules with the introduction of correlation amongst different items or different individuals. With such rules, the system will be granted the ability to determine what a person needs most currently, rather than giving recommendations based on the person's previous activity history in the system. One example will be that if someone is looking for an oven in the kitchen, but he or she was browsing a dishwasher 2 months ago, the system will begin to give

recommendations on oven or other cooking utensils, rather than kitchen cleaning utensils. Like mentioned before, a more complicated version of the association rule will give out recommendation with more complicated consideration and calculations. The recommendation system will be referencing different traits of a person or by viewing at a variety of items being browsed in the system by the user. One supplementary of the association recommendation system will be using dynamic analyzing to provide recommendations for future use when the user wished to need some resources that related to the current inquiry. An example would be a person who bought or browsed an oven today may be provided information of recommendation on oven recipe, or aluminum foil tomorrow.

## 9.3 Big Data Recommendation Engines in Education

As we have mentioned before, recommendation engines based on data mining are proving to be beneficial to almost all fields in our lives, and education is one of them.

The field in education that involves big data mining are often referred as learning analytics. It focuses on how big data mining could be utilized for teaching and learning purposes ranging from personalized teaching, learning, evaluations and assessments for individuals to providing data to decision makers of various levels of education (for example, a director of a department or a government official of education). Big data mining has provided benefits to many aspects of education such as teaching, learning, education leadership, adult education, special education, enrollment decision, talent education, etc. The new millennium has seen the rapid development of educational big data mining and the field is hunger for talents that possess not only profound understanding in educational theory, but also the capability to carry out statistics, research and evaluation in education[32].

As the examples mentions above, the educational recommendation systems have deep similarities with commercial recommendation systems as they both strive to introduces the user to their desired products or services. However, educational recommendation system could also provide interconnection between learners, their desired course, their personal traits and educational resources that could serve the learners to help them reaching maximum efficiency in learning and to reach their academic goals. These beneficial factors make the educational recommendation engines a state-of-the-art asset for students to excel in personalized online learning systems. In such system, learner's characteristics, track selection and knowledge gained in previous learning could all be quantized into values to serve as a filtering and weighing standard to learners in e-learning. As one can see, such system has great flexibility and are highly adaptive to different learners. In this way, the efficiency of learning is greatly enhanced, and students are more motivated in engaging in learning[31].

The history of the e-learning recommendation system could be traced back to computer assisted instruction systems, also known as CAI. One major concept called Time-shared, Interactive, Computer-Controlled, Information Television (TICCIT) was invented in the last 70s. This could be the cornerstone of nowadays educational recommendation engine. TICCIT is developed so that the learner

could have higher control in their own learning with the help of a mentor giving suggestions and advices from time to time. The education recommendation has met its rapid development afterward ever after the introduction of TICCIT as they could provide personalized advice and suggestions to learners according to their daily usage and browsing history of the system. Students could spend less time on looking for the education resources on their own or filtering out valued teaching and learning resources from a gigantic amount of information on the internet or within the e-learning system. In such way they could devote all their valuable time to learning, rather than being in a frustrated state without guidance[22].

As mention before, the recommendation system's ability to provide an accurate result relies on massive amount of data collected from individuals and their behavior on the e-learning website. In this way, e-learning websites with a considerable number of users could better contribute to the learning process of the recommendation engine. For instance, Massive Open Online Courses (MOOCs) could have hundreds or thousands of active users on the website, or even learning the same course at one time. In such way, the recommendation engine evolves quickly, and user could benefit from it. Moreover, online learning websites have a social learning ecosystem which have great resemblance to social media networks. This lays the groundwork for the recommendation system to make full utilization of its huge user database to provide more relevant courses for learners. It is worth mentioning that these e-learning systems with social element are more likely to be involved with informal or professional learning. An example would be info of a user in career development system will be put under comparison to his or her colleague's information to carry out a performance evaluation[16].

Also, when he or she wishes to visit some of the resources on the career training website, the system could filter out his or her colleague's recommendation, comment, rate of one course and then utilize algorithm to provide this user with resources not only capable of helping him or her reaching current goals, but also courses and information that may become useful in the future[21]. Lots of educational teaching and learning systems with data mining and recommendation feature are established on online learning systems that could be easily visited from a mobile phone or a tablet. Such convenience no doubt made collecting data at great ease and allows more users to participate in such process. This could be a beneficial cycle: the ever-growing user base allows the algorithm to be more accurate and provide more personalized learn guidelines, and such feature will not only attract more user but will also let remaining users to provide more data to the system.

Learning analytics could also find itself useful other than the scenario of teaching and learning systems. Data mining and recommendation engines could be also used in supporting students in daily learning. For instance, a system that feature in college application could use a recommendation engine to provide learning track for students to better prepare for a certain university's requirements. Such method could be also used with other kinds of online learning motivation techniques such as badges. Badges are like achievement system in which when a student accomplish certain goals, he or she will be award a badge. He or she could get to know the global percentage of student holding that badge and get

motivated in making more accomplishments[19].

One more application would be the student retention system, in which students' data are monitored and a baseline is set based on the overall performance of all the student within. If one student's performance is below par, the system will receive alert and will send support or intervention staff as soon as they can to help the student and prevent him or her from dropping the course. In addition, big data also provide new thinking on how to conduct the PBL learning process. For example, when learners are working in the virtual environment, tutors can monitor the contribution of students in a certain group. In this way, the tutor can quickly identify who is not contributing to the team and take certain measures to intervene the performance of this student. Moreover, in a conventional PBL environment, the timing of proving scaffolding and removing scaffolding is very hard to master. But with the help of big data, tutors can analyze the process of learning in a team and spot the time when the team make minimal progress[15]. In this way, the instructors and tutors can provide in-time support. And tutors can also spot the time when a team have sufficient knowledge and ability of accomplishing the task, so tutors can fade scaffolding. From the learners' perspective, big data give them space to try new ideas. Instead of having group discussions and debates of different ideas, students can also learn from what the data tells them and gain empirical experience. With big data technology, learners can also formulate more up-to-date solutions to a task[18].

Big data also provide powerful tools to instructional designers. With the help of data, instructional designers can label learners just as the way online retailers label customers. Then those labels will be put into different categorizes. This is very important for conducting learner analysis. Instructional designers will be able to see clearly the motivations they have, the prior knowledge they possess to determine the scope of learning. Also with such data, instructional designers can design proper strategies to motivate learners and increase the satisfaction rate. The learning style data will help the development of teaching strategies. Designers and subject-matter experts can integrate different ways of learning in one semester based course and let learners with different learning habits to collaborate to foster flexibility in learning.

## 10 WEB ANALYSIS

Web analysis is also an uprising branch that belongs to the learning analytics and being supported by big data mining. It focuses on how to collect and analyze data gained on websites or applications that needs to connect to internet before using. These kinds of data are often a result of user's activities on the internet. Web analytics are often utilized to boost the study and learning efficiency of students in a specific Learning Management System(LMS). It could also help the administrator of the website to monitor and support student's learning progress and to help oversee the functioning of the website[17].

Web analytics are also utilized by administrators to get a better understanding of what kinds of personal traits one user may carry and how would this one interacts with various function on the website. It could be utilized to make a prediction on what kind of educational

products or courses will be more welcomed by certain students and learners. After analyzing such data as how many people have visited one page and what kind of activities they are most likely to carry out, the administrator could be informed that what kind of needs one student possesses and how they can develop in the future to cater to their needs. For the education website owners, the web analysis can also be used as a mean to find out any hidden security risks online and could help them gaining evidence for court should an attack really happens.

In the world of academics, web analytics is also beneficial in helping the college to make strategic plans. For instance, with a growing number of traditional courses, tutoring services are going to be changed into their online version, web analytic will find out how to deploy these courses and servers better so that they could get maximum visit from those who are interested in them. Since many data and information are distributed on different websites, they call for the facilitation of web analytics to perform an integration to the scattered resources. Moreover, the educational corporations, both online and offline, could easily get to know how the traffic flow changes every day on their online learning systems[1].

## 10.1 Web Analytics Anatomy

Normally, the history of web analytics could be traced back to last 60s when scientists start to analyze web logs. These logs could be transaction or search types. The transaction type takes direct actions such as user's clicks, how long they have spent on one page into consideration while search type focuses more on the behavior on how the users carry out the searching activities.

Depending on the data provided by the servers, web analytics send small package of data (commonly known as cookies) to the user. Cookies will start collecting data and send them back to the server. Such process is called as server-side data collection. On the other hand, this kind of data collection would render itself not accurate. Internet service provider (ISP) provides IP address to users while user many set blockade to some cookies. To the contrary, client-side data collection is more flexible and can be more accurate. By implanting tags into the website being visited by the client, client-side data collection could carry out more versatile missions.

The word Human Computer Interaction(HCI) have been a buzzword nowadays and it have been embedded into our daily lives. Web analytics could also have utilized such different methods as interviews, questionnaires to establish more convincing reports. Key Performance Indicators (KPI) are set up to differentiate various kinds of web analytics. As of an example, one university that introduced with a new kind of LMS are facing difficulty because too much people are using its social features and it needs some backup support. The web based analysis will be performed to assess the resource the university possesses and evaluate the need to figure how to employ capable person to perform certain kinds of maintenance work as well. The KPI within could be able to indicate how many clicks can one user click before reaching the help page, how long will a user spend on the help page, how easy the help material could be comprehended, how visible are the various icons to the viewer and so on. Then the KPIs are collected and analyzed to compose a report[4].

## 10.2 Web Analytics and Education

In the field of education, web analytics are utilized to form reports that are driven by data to help such functions as facilitating students, managing staffs and supporting researches. Also, web analytics are used to figure out how well a student could perform, how would the student and online tutor would normally interact, how effective one course could be and how well the student is progressing in the course. One specific kind of web analysis is called as academic analytics. It would evaluate the overall performance on an online teaching website to provide information so that administrators could better make decisions.

Learning analytics, as mentioned before, focuses on the collection and analysis of data that have relation to the learners and the courses and learning materials. Learning analytics are also utilized in documenting students' overall learning efficiency in computer facilitated learning and could facilitate student to get accustomed to the online teaching and learning environment better. With big data gained and stored in the systems, learning analytics made many contributions to lots of fields that could help students reaching their academic success. For example, they could note down where the students are now in the middle of a course; if a student misses too much class, they could figure it out and send intervene staff quickly; assess different aspect of the Learning management system; give out help and facilitation that could cater to a student's need[5].

As one way to lead students to academic success, learning analytics could trace all students' activity and other behaviors in the online environment, with data collected via the student information system (SIS). In a class that is one hundred percent online, instructors could fully utilize learning analytic to carry out formative evaluation, which could help the teacher to learn about how the students are performing, how could they make modification to ongoing courses, and how he or she could demonstrate such course materials to the students. An example would be that the teacher could track how many time the student have entering the LMS in the allotted time and use it as evidence of attendance record. Also, the teacher could record how many clicks are carried out in one content page or during one course, or how long the student has spent in different sections of the course. All the data collected above could give out information on how the user behave and how the relations of learners and teaching and learning materials have been. With the deployment of web learning analytic in the LMS, the instructor could figure out abnormal activities of students and give out interventions that cater to the student's needs.

To boost the efficiency of the learning analysis system to the maximum level, learning analytic could also reach to qualitative data such as the discussions in students' forums, students' cooperative wiki pages, and many other social learning assets to form more persuasive and convincing data to website administrators. This requires natural language processing kit (NLTK) to perform semantic analysis so that these qualitative data could be better transcribed into data that would be better analyzed. These data gained could be utilized to perform some higher-level assessments, such as the creativity and critical thinking level of a student[7]. With different teaching and learning goals, KPI could help student, or make modifications of online course in the online learning website with the facilitation of quantitative and qualitative data. They can also

run course diagnose for learners and teachers. The KPI mentioned could be collected and analyzed with such techniques as students' characteristics and performance tracking, investigating a group of students with same traits, giving out content recommendations of learning materials on history activities and make prediction to future developments.

## 11 THE INTELLIGENT TUTORING SYSTEMS

The term intelligent tutoring system are used to describe a computer system that could act as a human mentor to some extent to facilitate a student in getting to understand and have firm understanding of the learning materials. Such system is often designed to make learning with a higher efficiency as well as providing inspiration to students while learning. It requires the support of big data and are considered one of the rising learning technologies in the field[26]. Taking a human mentor for example, he or she will be preparing for the learning material for the student first, then he or she will try the best to get the student motivated for learning. When the student is facing difficulties in learning, he or she will stand out and provide necessary guidance for them to overcome the barrier. Likewise, AI of the intelligent tutoring system could be evaluated and determined whether it could qualify as a human teacher. Such system need to negotiate and communicate with a student to get accustomed to newer conditions, and when a student make requests of learning materials or asks question on certain items, the system will adjust automatically to cater to student's need better. In a word, intelligent tutoring system is different from commonplace e-learning websites as it is more flexible and could provide more detailed education contents. This system could store gigantic amount of data, ready to respond to unique needs of students under different scenarios[12].

### 11.1 Anatomy of the ITS

Such system has provoked the wide interest amongst researcher and programmer thus many have devoted themselves into designing it, which makes one kind of such system greatly differs from another. It is worthwhile to notice that even these systems have distinctive design theories, they are share the similarities of the following elements: domain, learner, pedagogical and interaction model.

Domain model mainly answer the question on how to represent the core knowledge on the computer. It could be demonstrated as flow charts, diagrams, semantic networks, etc. It is mainly consisted of the fundamental logic, strategies and rules to solve ongoing questions. It is the logical core of the system, as it will provide assessment standard when the student's progress is going through evaluation. Moreover, it will also serve as a detector of abnormal behaviors[8].

Learner's model will be demonstrating the systematic evaluation on how well the student is going through one course, what kind of error the student will most likely to make, what kind of learning style the student prefers, what characteristic the student possesses, etc. Such information is collected via students' activities on the system. This kind of model is also utilized in self-regulated learning,

which relies little on the help of other human instructors.

Pedagogical model focuses more on using best teaching and learning strategies to the student according to the teaching environment. It will check on the student's learning progress, and give out appropriate information or facilitation accordingly.

Interaction model, also known as the interactive model are more like a translator between the system and the student. It will need to receive student's input and give out response that could meet the student's needs. Not only this model requires information on the learning material, it would also need information on the common sense of mankind. It was based on verbal texts but nowadays one could identify users' interaction from a variety of sources such as facial expressions, body temperature and moisture, minor gestures, etc.

These four models are all under the management of a database. Moreover, the models are designed under the guidance of different educational theories and uprising technologies. As for interactive model, it is based on various multimedia means. To better understand one student's input, NLTK is involved as well as voice recognition software. AIs are introduced to interactive model to automatically output text and voice messages. Capturing technologies are also involved to capture the student's facial expressions and body gestures. Researchers are also trying to bring virtual and augmented reality to the model. This model involves psychology related content as well as researchers are managing to deploy emotional detection technology to determine one's affection state as it might cast profound impact on the learning effect of an individual. When the internet haven't reached today's popularity, many of the ITS were installed on the PC and cannot get frequent upgrade. Due to the hardware limitations of PCs at that time, the function of such ITS is highly limited, as there was little storage space, and PC didn't have high processing speed at that time[9].

With the rapid development of the internet and PC, ITS have entered an new era as many calculation and data could be processed on the cloud. This have removed the blockade of those who with to be guided by ITS and it is making a growing number of learners benefit from ITS. Nowadays the needed learning materials could be searched and retrieved in no time thanks to ever-developing searching techniques. Moreover, more online wikis are being established which provided supplementary source of domain knowledge to help broaden the borders of domain models.

Learners have also witnessed the rapid development of mobile learning in the field of ITS. Wherever there is internet, learner could easily get in touch with ITS at every corner of the world. As smart cell phones and tablets became almost necessities of everyone, ITS have also taken a leap forward and keep absorbing the newest discoveries in such field as machine learning and big data mining. Learners nowadays could get authentic and quick feedback from ITS, which could be a great motivation to the process of learning.

### 11.2 Designing ITS

To design an ITS system, researchers have to follow certain procedures, which have certain resemblance with designing a learning management system, or a teaching and learning software. It is commonly agreed that such process take place in four steps: Needs

assessment to carry out the anatomy of learner goals and discussion with the instructor and course material designer for the course; Cognitive task analysis to start building models mentioned before, preparing to tackle any issues in the developmental process; Initial mentor implementation to set up the ecosystem of the ITS and to provide learning facilitation; Evaluation to start trial runs of the ITS and to testify the overall steadiness and robustness of the ITS and give out a holistic assessment to the system.

### 11.3 Applying ITS

One of the most outstanding feature of ITS is that it possesses the capability to give out immediate response to students' needs without a human teacher. Moreover, it can also give timely support, choosing different learning goals for students with different demands, giving individualized coaching and provide mental reinforcement. As a consequence, ITS are more likely to be deployed at institutions, army camps, and business where tutoring and mentoring is required in training yet lacks enough human tutor. Ergo, it have a wide spectrum of application, ranging from kindergarten education, to training on jobs, and even lifelong learning. Researchers have profound interest in studying the efficiency and other benefits of ITS. Such aspects of students as how well they could comprehend the course materials, how eager are they when learning about new contents, how much would they devote themselves into learning and how satisfied they will be after the learning are all taken into consideration. They will even arrange human tutoring session to compare the overall efficiency between human and computer tutors. Some researchers have found out that there is only Little difference between the effect of human tutoring and machine tutoring[10].

As the developers have reached the goal of giving response immediately and provide escalated tutoring techniques, they are facing new challenges now. Due to the complexity of such system, ITS is not economy-friendly to design and deploy. As a result, researchers and developers are researching and developing means to make deploying these systems at a lower cost.

### 11.4 Envisioning ITS

As mentioned before, the most vital feature of ITS is it does not require extra human tutor to give help to the student. What is more, it could also generate and comprehend natural language for better communication between the machine and the learners. There are many undiscovered areas for researchers to venture in as the recognition rate of the system are still in a moderate level and still have rooms for development. Also, the natural language output give by the ITS sometimes are not considered authentic enough for students to understand. Researchers are also calling for the research on the identification of students' affection state so that the dialogue may change to different mood the student is in accordingly. The system should also be capable of know how the student will be most motivated, thus planning for motivation strategies.

The researchers also have the ambition to upgrade the ITS from a system to an environment. It could adapt to more kinds of learners, providing more reliable content and support to learners, and have greater flexibility in the tutoring process. The most advanced ITS

could still only function in questions that have clear boundaries and finite solutions. It is all researchers' hope that in the future the ITS will be able to support student with question that have open answers[11].

## 12 CONCLUSION

For centuries, the learning style of countries around the world remains similar. The teachers are served as the center of information. Students go to school to acquire information they otherwise do not know if they just stay at home. That is the reason why behaviorism was proposed as the main theory of learning. However, with the development of science of technology, people have more tools of getting information such as radio, television and movies. Such development pushed educators to revise their educational strategies in order to make education attractive. Then cognitive theory came into being and provided guidance of how to facilitate the process of learning. However, with the increase of publishing of books and the development of affordable and high speed internet, teachers can no longer serve as people who provide information to students in the fast-changing world. Therefore the strategy of teaching must again change the suit the world. The shift is to foster students ability to solve ill-defined problems through collaboration with minimal guidance from instructors and develop meta-cognitive skills. Data mining and recommendation engine have proven their importance nowadays and will continue to shine and make more impact in the foreseeable future in the field of education. The specific field of learning analytic will continue to make more contribution to online learning. However, we must take the ethical use of big data into consideration and make sure we maximize the benefit of big data in education while preventing misusing the data and protect individual's privacy at all costs. Another issue might be the ever-complex algorithms and codes will be a challenge to education specialists and school or learning website admins while the programmer may have limited knowledge in education. However, with the rapid development of educational recommendation system, many new job opportunities will be created, thus encouraging specialists to carry out interdisciplinary research and there will be a growing number of talents that excel both in education theories and programming. It also calls for the tight collaboration between education expert and programmers to make sure that the ever growing education recommendation system backed by big data mining will lead countless of learner to their academic success. The potential of big data on education is still not clear. Although big data have been employed in commerce, healthcare, artificial intelligence and other industries, educators are still waiting to see its implication on learning. However, we can predict big data will bring positive changes to learning as a whole and provide new perspective to instructional design.

## A CONCLUSION OF ROLES IN THE TERM PAPER

In this term paper my partner Weipeng Yang and I participated in the discussion of the general topic of the paper. We finalized the topic through a meeting. Since we are all students of the Instructional System Technology department at the school of education, we reached an agreement that the topic should be how big data can

influence instructional design.

Then in the following meetings we had , we generally came up the the structure of the paper. Instructional design is a subject of education. However, instructional design itself is still a broad topic. Therefore, we want to put our focus on Problem-based learning which is an important learning strategy. In addition, we also took our audience into consideration. The potential readers of the paper are not necessarily in the field of instructional design, so we thought it is important to introduce this field of study, the development of instructional design first. And then we will focus on Problem-based learning and give a few examples.

And because we are not in the field of big data and this field is really strange to us, we wanted to summarize topics involved in big data and then present how big data will influence instructional design.

In this term paper, my responsibility was to focus on the instructional design part and Weipeng was in charge of the big data part. But we also participated in each other's work to keep the group go smoothly.

## REFERENCES

- [1] S Aher and L Lobo. 2013. Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. *Knowledge-Based Systems* (2013).
- [2] Bob Burgess. 2011. The Educational Theory of Socrates. (2011). <http://www.newfoundations.com/GALLERY/Socrates.html>
- [3] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. *MIS QUARTERLY* 36, NO.4 (Dec. 2012), 1168–1169.
- [4] W Chughtai, A Selama, and I Ghani. 2013. Short systematic review on e-learning recommender systems. *Journal of Theoretical & Applied Information Technology* (2013).
- [5] B Clifton. 2008. (2008). <http://www.ga-experts.com/web-data-sources.pdf>
- [6] R S Donner and H Bickley. 1993. Problem-based learning in American medical education: an overview. *Bulletin of the Medical Library Association* 81, 3 (1993), 294–298.
- [7] P Dwivedi and K Bharadwaj. 2013. Effective trust-aware e-learning recommender system based on learning styles and knowledge levels. *Educational Technology & Society* (2013).
- [8] R Ferguson. 2012. Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning* (2012).
- [9] J Gray, T Boyle, and C Smith. 1998. A constructivist learning environment implemented in Java. In *Proceedings of the 6th Annual Conference on the Teaching of Computing and the 3rd Annual Conference on Integrating Technology Into Computer Science Education: Changing the delivery of computer science education* (pp. 94f97).
- [10] T Gunz and M Hollingsworth. 2013. The implementation and assessment of a shared 21st century learning vision: A districtbased approach. *Journal of Research on Technology in Education* (2013).
- [11] M Hofer and N Grandgenett. 2012. TPACK development in teacher education: A longitudinal study of preservice teachers in a secondary M.A.Ed. Program. *Journal of Research on Technology in Education* (2012).
- [12] B Jansen. 2009. *Understanding user-web interactions via web analytics*. San Rafael, CA: Morgan & Claypool.
- [13] A. Januszewski and M Molenda. 2008. *Definition. In Educational Technology: A Definition with Commentary*. New York: Lawrence Erlbaum Associates, Chapter 1, 1–14.
- [14] Dan Kerns. 2013. 10 Ways Big Data is Changing K-12 Education. (2013). <http://www.dreambox.com/blog/10-ways-big-data-changing-k-12-education-2>
- [15] M Kim and E Lee. 2013. A multidimensional analysis tool for visualizing online interactions. *Journal of Educational Technology & Society* (2013).
- [16] K Kingsley and J Brinkerhoff. 2011. Web 2.0 tools for authentic instruction, learning and assessment. *Social Studies and the Young Learner* (2011).
- [17] Tom Krawiec. 2017. The Amazon Recommendations Secret to Selling More Online. (2017). <http://rejoiner.com/resources/amazon-recommendations-secret-selling-online/>
- [18] Harlan M. Krumholz. 2014. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff (Millwood)* (2014).
- [19] C Lankshear and M Knobel. 2006. *New literacies: Everyday practices & classroom learning* (2nd ed.). New York, NY: Open University Press.
- [20] How Big Data Will Boost Learning and Teaching in Higher Education. 2016. Cogbooks. (2016). <https://www.cogbooks.com/2016/10/05/big-data-will-boost-learning-teaching-higher-education/>
- [21] J Lucas, S Segarra, and M Moreno. 2012. Making use of associative classifiers in order to alleviate typical drawbacks in recommender systems. *Expert Systems With Applications* (2012).
- [22] R Maloy, R Verock-O'Loughlin, S Edwards, and B Woolf. 2014. *Transforming learning with new technologies* (2nd ed.). Boston, MA: Pearson.
- [23] DI Mansur, SR Kayastha, R Makaju, and M Dongol. 2012. Problem Based Learning in Medical Education. *Kathmandu University Medical Journal* 10, 4 (2012), 78–82.
- [24] Bernard Marr. 2015. How Big Data Is Changing Healthcare. (2015). <https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#86305c528730>
- [25] M. Molenda and E Boling. 2008. *In Educational Technology: A Definition with Commentary*. New York: Lawrence Erlbaum Associates, Chapter 4, 81–83.
- [26] Office of Educational Technology. 2013. Expanding evidence approaches for learning in a digital world. (2013). <http://tech.ed.gov/files/2013/02/Expanding-Evidence-Approaches.pdf>
- [27] Kaushik Pal. 2016. What is the influence of Big Data in Medicine? (2016). <https://www.kdnuggets.com/2016/03/influence-big-data-medicine.html>
- [28] Evelyn Sakakida RaWski. 1979. Education and Popular Literacy in China Michigan. (1979).
- [29] R. A. Reiser and J. V. Dempsey. 2012. *Trends and issues in instructional design and Technology*: (3 ed.). Boston, MA: Pearson Education, Inc., Chapter What field did you say you were in? Defining and naming our field, 1–7.
- [30] Jeffrey Riegel. 2013. Confucius. (2013). <https://plato.stanford.edu/entries/confucius/>
- [31] C Romero, S Ventura, and E Garcia. 2008. Data mining in course management systems: Moodle cases study and tutorial. *Computers & Education* (2008).
- [32] C Romero, S Ventura, A Zafra, and P de Bra. 2009. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computer and Education* (2009).
- [33] J Rountree, N Rountree, A Robins, and R Hannah. 2005. Observations of student competency in a CS1 course. In *Proceedings of the 7th Australasian Conference on Computing Education: Vol. 42*.
- [34] P Saettler. 1990c. *Behaviorism and educational technology: 1950 - 1980*. Englewood, CO: Libraries Unlimited, Chapter 10, 293.
- [35] J. R. Savery and T. M Duffy. 2001. *Problem-based learning: An instructional model and its constructivist framework*. Technical Report 16. Indiana University Bloomington.
- [36] Eric Schadt. [n. d.]. The role of big data in medicine. ([n. d.]). <https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-role-of-big-data-in-medicine>
- [37] J Schafer. 2005. *The application of data mining to recommender systems*. Hershey, PA: Idea Group.
- [38] L Schrum and B Levin. 2009. *Leading 21st century schools: Harnessing technology for engagement and achievement*. Thousand Oaks, CA: Corwin.
- [39] S Shum and R Ferguson. 2012. Social learning analytics. *Journal of Educational Technology & Society* (2012).
- [40] G Siemens. 2013. Learning analytics: The emergence of a discipline. *American Behavioral Scientist* (2013).
- [41] K. H. Silber and W. R Foshay. 2006. *Handbook of human performance technology*. San Francisco: Pfeiffer, Chapter Designing instructional strategies: A cognitive perspective, 371.
- [42] B.F. Skinner. 1954. The science of learning and the art of teaching. *Harvard Educational Review* (1954).
- [43] Wikipedia. 2017. Education. (2017). <https://en.wikipedia.org/wiki/Education>
- [44] B. G Wilson. 2012. *Trends and issues in instructional design and technology* (3 ed.). Boston, MA: Pearson Education, Chapter Constructivism in practical and historical context, 45.
- [45] P Winoto, T Tang, and G McCalla. 2012. Contexts in a paper recommendation system with collaborative filtering. *International Review of Research in Open and Distance Learning* (2012).
- [46] Diana F Wood. 2003. ABC of learning and teaching in medicine. *Clinical review* 326 (2003), 328–329.

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 "Wikipedia2017" isn't style-file defined  
--line 94 of file report.bib  
Warning--entry type for "Riegel2013" isn't style-file defined  
--line 101 of file report.bib  
Warning--entry type for "Burgess2011" isn't style-file defined  
--line 109 of file report.bib  
Warning--entry type for "Schadt" isn't style-file defined  
--line 146 of file report.bib  
Warning--entry type for "Pal2016" isn't style-file defined  
--line 152 of file report.bib  
Repeated entry--line 184 of file report.bib  
: @article{Wikipedia2017  
:  
,  
I'm skipping whatever remains of this entry  
Warning--entry type for "Marr2015" isn't style-file defined  
--line 191 of file report.bib  
Warning--entry type for "clifton2008" isn't style-file defined  
--line 198 of file report.bib  
Warning--entry type for "Office2013" isn't style-file defined  
--line 330 of file report.bib  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing--line 3085 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing--line 3131 of file ACM-Reference-Format.bst

Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end while executing---line 3131 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001 while executing---line 3131 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001 while executing---line 3131 of file ACM-Reference-Format.bst  
Warning--no number and no volume in Aher2013  
Warning--page numbers missing in both pages and numpages fields in Aher2013  
Warning--no number and no volume in Chughtai2013  
Warning--page numbers missing in both pages and numpages fields in Chughtai2013  
Warning--no number and no volume in Dwivedi2013  
Warning--page numbers missing in both pages and numpages fields in Dwivedi2013  
Warning--no number and no volume in Ferguson2012  
Warning--page numbers missing in both pages and numpages fields in Ferguson2012  
Warning--empty publisher in Gray1998  
Warning--empty address in Gray1998  
Warning--page numbers missing in both pages and numpages fields in Gray1998  
Warning--no number and no volume in Gunn2013  
Warning--page numbers missing in both pages and numpages fields in Gunn2013  
Warning--no number and no volume in Hofer2012  
Warning--page numbers missing in both pages and numpages fields in Hofer2012  
Warning--empty address in Jansen2009  
Warning--empty address in Januszewski2008  
Warning--no journal in Kerns2013  
Warning--no number and no volume in Kerns2013  
Warning--page numbers missing in both pages and numpages fields in Kerns2013  
Warning--no number and no volume in Kim2013  
Warning--page numbers missing in both pages and numpages fields in Kim2013  
Warning--no number and no volume in Kingsley2011  
Warning--page numbers missing in both pages and numpages fields in Kingsley2011  
Warning--no journal in Krawiec2017  
Warning--no number and no volume in Krawiec2017  
Warning--page numbers missing in both pages and numpages fields in Krawiec2017  
Warning--no number and no volume in Krumholz2014  
Warning--page numbers missing in both pages and numpages fields in Krumholz2014  
Warning--empty address in lankshear2006  
Warning--no journal in Learning2016  
Warning--no number and no volume in Learning2016  
Warning--page numbers missing in both pages and numpages fields in Learning2016  
Warning--no number and no volume in lucas2012  
Warning--page numbers missing in both pages and numpages fields in lucas2012  
Warning--empty address in Maloy2014  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end while executing---line 3229 of file ACM-Reference-Format.bst

Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--empty address in Molenda2008  
Warning--no journal in RaWski1979  
Warning--no number and no volume in RaWski1979  
Warning--page numbers missing in both pages and numpages fields in RaWski1979  
Warning--empty address in Reiser2012  
Warning--no number and no volume in Romero2008  
Warning--page numbers missing in both pages and numpages fields in Romero2008  
Warning--no number and no volume in Romero2009  
Warning--page numbers missing in both pages and numpages fields in Romero2009  
Warning--empty publisher in Rountree2005  
Warning--empty address in Rountree2005  
Warning--page numbers missing in both pages and numpages fields in Rountree2005  
Warning--empty address in Saettler1990c  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--empty year in Schadt  
Warning--can't use both author and editor fields in Schafer2005  
Warning--empty address in Schafer2005  
Warning--empty address in Schrum2009  
Warning--no number and no volume in Shum2012  
Warning--page numbers missing in both pages and numpages fields in Shum2012  
Warning--no number and no volume in Siemens2013  
Warning--page numbers missing in both pages and numpages fields in Siemens2013  
Warning--empty address in Silber2006  
Warning--no number and no volume in Skinner1954  
Warning--page numbers missing in both pages and numpages fields in Skinner1954  
Warning--empty address in Wilson2012  
Warning--no number and no volume in Winoto2012  
Warning--page numbers missing in both pages and numpages fields in Winoto2012  
(There were 21 error messages)  
make[2]: \*\*\* [bibtex] Error 2

latex report

=====

[2017-12-10 13.49.57] 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 1.2s.
```

---

## Compliance Report

---

```
name: Yang Weipeng
hid: 236
paper1: Oct 22 17 - 100%
paper2: Nov 7 17 - 100%
project: Dec 4 17 - 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
12
wc 236 project 12 12097 report.tex
wc 236 project 12 12584 report.pdf
wc 236 project 12 1481 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
12: \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
```

---

```
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: figure and above may be used improperly

190: As one way to lead students to academic success, learning analytics could trace all students' activity and other behaviors in the online environment, with data collected via the student information system (SIS). In a class that is one hundred percent online, instructors could fully utilize learning analytic to carry out formative evaluation, which could help the teacher to learn about how the students are performing, how could they make modification to ongoing courses, and how he or she could demonstrate such course materials to the students. An example would be that the teacher could track how many time the student have entering the LMS in the allotted time and use it as evidence of attendance record. Also, the teacher could record how many clicks are carried out in one content page or during one course, or how long the student has spent in different sections of the course. All the data collected above could give out information on how the user behave and how the relations of learners and teaching and learning materials have been. With the deployment of web learning analytic in the LMS, the instructor could figure out abnormal activities of students and give out interventions that cater to the students needs. \\

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--entry type for "Wikipedia2017" isn't style-file defined  
--line 94 of file report.bib  
Warning--entry type for "Riegel2013" isn't style-file defined  
--line 101 of file report.bib

```
Warning--entry type for "Burgess2011" isn't style-file defined
--line 109 of file report.bib
Warning--entry type for "Schadt" isn't style-file defined
--line 146 of file report.bib
Warning--entry type for "Pal2016" isn't style-file defined
--line 152 of file report.bib
Repeated entry---line 184 of file report.bib
: @article{Wikipedia2017
:
I'm skipping whatever remains of this entry
Warning--entry type for "Marr2015" isn't style-file defined
--line 191 of file report.bib
Warning--entry type for "clifton2008" isn't style-file defined
--line 198 of file report.bib
Warning--entry type for "Office2013" isn't style-file defined
--line 330 of file report.bib
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001
while executing---line 3131 of file ACM-Reference-Format.bst
Warning--no number and no volume in Aher2013
Warning--page numbers missing in both pages and numpages fields in Aher2013
Warning--no number and no volume in Chughtai2013
Warning--page numbers missing in both pages and numpages fields in Chughtai2013
Warning--no number and no volume in Dwivedi2013
Warning--page numbers missing in both pages and numpages fields in Dwivedi2013
```

Warning--no number and no volume in Ferguson2012  
Warning--page numbers missing in both pages and numpages fields in Ferguson2012  
Warning--empty publisher in Gray1998  
Warning--empty address in Gray1998  
Warning--page numbers missing in both pages and numpages fields in Gray1998  
Warning--no number and no volume in Gunn2013  
Warning--page numbers missing in both pages and numpages fields in Gunn2013  
Warning--no number and no volume in Hofer2012  
Warning--page numbers missing in both pages and numpages fields in Hofer2012  
Warning--empty address in Jansen2009  
Warning--empty address in Januszewski2008  
Warning--no journal in Kerns2013  
Warning--no number and no volume in Kerns2013  
Warning--page numbers missing in both pages and numpages fields in Kerns2013  
Warning--no number and no volume in Kim2013  
Warning--page numbers missing in both pages and numpages fields in Kim2013  
Warning--no number and no volume in Kingsley2011  
Warning--page numbers missing in both pages and numpages fields in Kingsley2011  
Warning--no journal in Krawiec2017  
Warning--no number and no volume in Krawiec2017  
Warning--page numbers missing in both pages and numpages fields in Krawiec2017  
Warning--no number and no volume in Krumholz2014  
Warning--page numbers missing in both pages and numpages fields in Krumholz2014  
Warning--empty address in lankshear2006  
Warning--no journal in Learning2016  
Warning--no number and no volume in Learning2016  
Warning--page numbers missing in both pages and numpages fields in Learning2016  
Warning--no number and no volume in lucas2012  
Warning--page numbers missing in both pages and numpages fields in lucas2012  
Warning--empty address in Maloy2014  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 3 in "DI Mansur, and SR Kayastha and R Makaju, and M Dongol" has a comma at the end  
while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--empty address in Molenda2008  
Warning--no journal in RaWski1979  
Warning--no number and no volume in RaWski1979  
Warning--page numbers missing in both pages and numpages fields in RaWski1979  
Warning--empty address in Reiser2012  
Warning--no number and no volume in Romero2008

Warning--page numbers missing in both pages and numpages fields in Romero2008  
Warning--no number and no volume in Romero2009  
Warning--page numbers missing in both pages and numpages fields in Romero2009  
Warning--empty publisher in Rountree2005  
Warning--empty address in Rountree2005  
Warning--page numbers missing in both pages and numpages fields in Rountree2005  
Warning--empty address in Saettler1990c  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Savery, J. R., and Duffy, T. M" has a comma at the end for entry Savery2001  
while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--empty year in Schadt  
Warning--can't use both author and editor fields in Schafer2005  
Warning--empty address in Schafer2005  
Warning--empty address in Schrum2009  
Warning--no number and no volume in Shum2012  
Warning--page numbers missing in both pages and numpages fields in Shum2012  
Warning--no number and no volume in Siemens2013  
Warning--page numbers missing in both pages and numpages fields in Siemens2013  
Warning--empty address in Silber2006  
Warning--no number and no volume in Skinner1954  
Warning--page numbers missing in both pages and numpages fields in Skinner1954  
Warning--empty address in Wilson2012  
Warning--no number and no volume in Winoto2012  
Warning--page numbers missing in both pages and numpages fields in Winoto2012  
(There were 21 error messages)

#### bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 8220  
non ascii found 8220

```
non ascii found 8221
non ascii found 8217
non ascii found 8220
non ascii found 8221
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 Monitoring Outdoor Air Quality

Janaki Mudvari Khatiwada  
Indiana University, Bloomington  
P.O. Box 1212  
Bloomington, Indiana 43017-6221  
jmudvari@iu.edu

## ABSTRACT

Outdoor air pollution is one of the risk factors of public health. Air pollution adds burden to public health. Both developing and developed world use new technology and expertise to monitor outdoor air quality. United States Environmental Protection Agency (USEPA) collects outdoor air quality data from state, local and tribal agencies through outdoor air quality monitors across the country. The data get collected into the Air Quality System (AQS) database. This data can be used for variety of purposes such as education, research and regulatory. Data from this data-mart is available for different time-series like hourly, daily, weekly, monthly and yearly. It gives us a real picture of outdoor air quality and measurements of pollutants present in the air in a particular time period. The data can be used for comparing air quality among different regions, raise awareness to general public so that they can play a role in reducing household air pollutants, to see the trend of air pollutants at different time periods in a day or a season, it can also be combined with emissions data for comparative study. Data analysis of pollutants help identify pollution hot spots. Since air quality is vital for public health and environmental health, air quality monitoring possesses great significance from public health perspective. It is worth looking at simple statistical values and level of air quality index value for the pollutants described as criteria pollutants as described by World Health Organizations. Simple statistical calculations in bigger datasets help understand the extent and source of problem. It helps in comparing past statistics with the present so helps in evaluations of action being taken during past periods. Looking at the overall mean value of criteria pollutants of 2016 and 2006 reveals improved air quality level to some extent for all core-based statistical areas. But the mean value of carbon monoxide has significantly increased over the ten years period.

## KEYWORDS

i523, hid330, Outdoor Air Quality, Big Data, Air Quality Index

## 1 INTRODUCTION

Outdoor air is a valuable natural resource that is vital to the health and existence of human beings and other forms of life. The outdoor air not only has clean air but has presence of various pollutants. Several health research have revealed that air pollutants are contributing factors for lung cancer, cardiovascular disease, acute and chronic respiratory conditions. World Health Organization (WHO) in 2013 has assessed that air pollution is carcinogenic to humans [14]. "In 2012 WHO estimated that 72 percent of outdoor air pollution-related premature deaths were due to ischaemic heart disease and strokes" [14]. Being aware of this fact, governments along with the scientists and the environmentalists help make policies to

combat air pollution. Each country has set their own standards for outdoor air quality to protect their citizen's health. Every nation's standards depend upon their economic, cultural, social and political needs. "The United States enacted its Clean Air Act (CAA) in 1970 and was amended in 1990 as a way to set stage for combating air pollution challenges" [1]. Since then, the country has made a lot of progress in improving air quality while sustaining a constant economic growth. After the enactment of CAA significant progress has been made in improving the outdoor air quality, reducing emissions levels from vehicles and power-plants. Over the period of 1990 and 2015, "national concentrations of air pollutants improved 85 percent for lead, 84 percent for carbon monoxide, 67 percent for sulfur dioxide (1-hour), 60 percent for nitrogen dioxide (annual), and 3 percent for ozone" [1]. Particulate matters "concentrations (24-hour) improved 37 percent and coarse particle concentrations (24-hour) improved 69 percent" between 2000 and 2015 [1]. Today, United States, European nations, India, China and other developing countries monitor outdoor air quality and use the collected data for identifying the particles present in the air, their contribution to various health problems, their sources, health research and also to find out the solutions to minimize their production level.

Thousands of air quality monitors are placed across the united States including US Virgin Islands and Puerto Rico. They are stationed based upon the significance of air quality effects on health. These monitors stream outdoor air quality data to a national air database system called Air Quality System (AQS). As a result big outdoor air data is generated constantly everyday. Air Quality System database is a national database where state, local and tribal agencies submit all of the data collected from thousands of air quality monitors across the United States. These huge databases are easily accessible in EPAs air data website via AQS. Besides air data, AQS database system has weather and emissions data. Emissions data provides data from vehicular, industrial and powerplants emissions records. Weather plays an important role in the quality of outdoor air. For example, high wind may disperse concentration of chemical particles. AQS database also called AQS Data Mart, has summary of yearly air quality data since the year of 1957. These data give an understanding of outdoor air quality and different particles present in the air and their sources. Source of air particles can be natural or human generated. Pollen, smoke from wildfires, mold, dust are some of the natural air pollutants. Similarly, emissions from power-plants, industries and vehicles, different substances and solutions that human have generated for various purpose are human generated air pollutants.

To set standard for air quality, Air Quality Guidelines was published by WHO in 1987 and have been revised in 1997 [11]. WHO guidelines set an international standard for air quality based on which countries around the world set their own standards to achieve

the goal set by WHO. Nitrous Oxide (NO<sub>2</sub>), Sulphur dioxide (SO<sub>2</sub>), Carbon Monoxide (CO), ground level Ozone, Particulate Matter (PM) among others, are some of the common hazardous air pollutants. Particulate matters are categorized into two categories, PM2.5 and PM10, based on the size of fine particle. Based on the value of Air Quality Index (AQI), USEPA has classified Outdoor air quality, AQI level as ‘Good’, ‘Moderate’, ‘Unhealthy for sensitive Groups’, ‘Unhealthy’, ‘Very Unhealthy’ and ‘Hazardous’ [7]. The AQI value range from 0 to 500. The agency has assigned colors (‘Green’, ‘Yellow’, ‘Orange’, ‘Red’, ‘Purple’ and ‘Maroon’ respectively) to each of the air quality categories [7]. It is shown in figure 2.

## 2 BIG DATA AND OUTDOOR AIR QUALITY

In US there are about 4,000 outdoor air quality monitors operated by state environmental agencies [6]. They constantly collect air data on harmful suspended particles present in the air and send them to a national database center which is AQS database. EPA has air quality database from last 27 years from around the states [2]. The size of EPA’s database is 25 GB. The data contains valuable information about the concentrations of different air pollutants in different time series; hourly, daily, weekly and yearly. Besides air quality data, AQS database also contains emissions data and weather data which are vital to the outdoor air air quality. Emissions data is basically data from vehicular and industrial emissions. They help to understand the source of different air pollutants and their role in air quality as well as they can be used for furthering research in limiting their emissions. Yearly summary data of AQS Data Mart can also be used to see the progress made in reducing harmful air pollutants over the years. For example, EPA reported that emission of SO<sub>2</sub> has reduced by 73 percent from 1990 to 2011 which is resulted primarily from electric utilities. Study of Emissions and air quality data gives an insight into the source of air pollutants and scientists can use these data to build a better industrial and vehicular models that reduces the emissions of pollutants. Similarly policy makers set vehicle emissions standards and industrial waste management.

India and China, two bigger economies in the world are battling worst air pollution. In recent years IBM is doing collaborative work with local authorities to combat air pollution in cities like Delhi, Beijing and Johannesburg by providing its data analysis platform called ‘Green Horizons’ [3]. The platform uses machine learning tools to analyze past weather forecasts data along with near real time data from optical sensors, air quality monitors and satellites to understand past forecasting models and build a better prediction models for future forecasts [3]. This prediction model helped Beijing enforce air quality control measures on traffic, construction and industry.

Weather directly affects outdoor air quality. For example in a windy day PMs can easily be spread in a neighboring regions and high temperatures increases ground level ozone [3]. Similarly, another example of relationship between big data analysis and air pollution is use of Microsoft’s tools in incorporating Beijing’s outdoor air data collected from conventional monitors along with data from “environmental monitoring stations, traffic systems, weather satellites, topographic maps, economic data, and even social media” [4].

Furthermore, OpenAQ is another platform besides EPA–AQS repository, which holds and hourly updates near–live air quality data from around the world. It claims to have “collected 133,494,377 air quality measurements from 8,054 locations from 47 countries. Data are aggregated from 98 government level and research-grade sources” [2]. The platform helps general public identify global hotspots for poor air quality and allow to have a look at the outdoor air quality where they live [2]. There is another open forum site called ‘Air–Now International’ where users from around the world can participate in sharing and information about air quality data. “It is an international version of USEPA’s air now system”[8]. Big volume of EPA’s data can be combined with another big data, census data to find out the portion of population breathing polluted air. This gives an understanding of health effects among general public. Since, monitoring stations generate big volume of air pollution data and regularly stream into, EPA’s open source, air database, any concerned individual can access live raw data from the website to find out about the quality of air they are breathing.

## 3 AIR POLLUTANTS

EPA has prioritized six major air pollutants that are commonly found all over the US. They pose significant threat to public health and environment. They are called ‘Criteria Pollutants’ and they are ground level ozone, fine particles or particulate matter (PM2.5 and PM10), nitrogen dioxide, sulphur dioxide, lead and carbon monoxide [1]. WHO has set the guidelines for each of these pollutants as shown in figure1.

PM<sub>2.5</sub> are particles less than or equal to 2.5 micrometers in diameter while PM 10 are particles less than or equal to 10 micrometers. “Sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust and water are the main components of PM ” [14]. These components combine with each other to form variety of mixtures in the air and can easily enter our lungs. Longer exposure to these substances increases the risk of lung cancer and cardiovascular disease [14].

Data on emissions from powerplants, industries and motor vehicles shows that emitted pollutants like various volatile substances and various forms of nitrogen oxides (NO<sub>2</sub>) are responsible for the formation of ground level ozone. Chemical reactions between these substances create ground level ozone directly in the air [1]. Chest pain, coughing, throat irritation and inflammation are common problems caused by ozone air pollution. The main source of NO<sub>2</sub> is emissions from heating, power generation and engines in vehicles and ships. SO<sub>2</sub> is another air pollutant produced mainly from burning of fossil fuels. Volcano is a natural resource so<sub>2</sub> release in the outdoor air. Longer exposure to this pollutant causes inflammation of respiratory tract [14]. The data on these pollutants have been regularly analyzed to see their trend. They have also been used in health research to have an understanding of their impact on people’s health. Keeping track of problems and source of problems help us in keeping problem at check. Some air pollutants are characterized as hazardous or toxic air pollutants. Some of the examples include benzene, cadmium, mercury, lead and asbestos.

## 4 HEALTH HAZARDS

Air pollutants, such as hazardous air particles can easily reach our lungs when we breathe. The effects are itchy, irritated throat, nose and inflammation of respiratory tract. Pollutants such as PM 20 block our airtubes. These pollutants badly affects people with asthma and bronchitis. WHO had estimated that in 2012, 3 million premature deaths worldwide due to outdoor air pollution, particularly due to exposure to particulate matter of 10 microns or less [14]. And in 2014 WHO has reported 7 million premature deaths worldwide [14]. Other pollutants such as lead, pesticides, arsenic also called as toxic pollutants are carcinogenic hence are responsible for lung cancer which is one of premature deaths. Carbon monoxide a very common air pollutant generated by combustion has been called a silent killer. Its health effects include nausea, vomiting and reduced neuro and cardiovascular behavior as it blocks oxygen transfer inside the body thereby might lead to death without knowing the real cause of death. Exposure to higher level of ground level ozone have serious health issues, while it affects people with asthma and bronchitis, other groups of people also experience coughing, shortness of breath, eventually inflammation of airways and development of chronic obstructive pulmonary disease [9].

Sulphur dioxide (so<sub>2</sub>) is a highly poisonous gas present in ambient air. It is a byproduct of burning of sulphur or product containing sulphur. Its main source is burning of fossil fuels especially in the powerplants and other industrial facilities. This pollutant can harm the environment by causing acid rain [10]. It harms human health by causing breathing difficulty and coughing. It combines with other particle pollutants present in the air causing haze.

## 5 AIR QUALITY INDEX

AQI is the index, for five major air pollutants discussed above, calculated by special formula developed by EPA. EPA uses its own formula to convert daily concentrations of measurement of each pollutant into AQI value of each pollutant [7]. Among all the highest AQI value is reported as the daily AQI value for that day [7]. Generally AQI 100 is the acceptable index set by EPA to protect public's health and it ranges from 0 to 500 [7]. The higher the value of AQI, greater is the pollution level and greater is the health risk. Based on hourly data collection from air quality monitors, stakeholders can constantly monitor AQI value in their cities or respective location. So weather channels in different media outlets such as local radio, television stations and newspapers also report about AQI index in order to inform general public about air quality in their area. Figure 2 shows the AQI classification for each pollutant as recommended by WHO and implemented by U. S. EPA. EPA is requires to report any AQI value greater than 100 specifically in larger cities with population more than 350,000 [7].

## 6 OUTDOOR AIR QUALITY MONITORING STATIONS

Outdoor(Ambient) air quality monitors are specified based on the significance of monitoring a particular pollutant [8]. The purpose might be to protect public health or environment in a densely populated areas. They might be stationed nearby, schools, hospitals, parks and recreational areas. While they are operated by several different agencies they are regulated by U. S. EPA. According to

EPA these stations should meet all the requirements for designs and operations as regulated by EPA themselves. These stations not only provide data on air quality they help in evaluating the effectiveness of programs and policies on emissions control.

## 7 WHO GUIDELINES AND CLEAN AIR ACT

WHO guidelines for air quality is applied worldwide. This guidelines was revised in 2005 [14]. The guidelines set standards for different air pollutants. According to the guidelines which is based on scientific evidence WHO has set standards for Ozone (o<sub>3</sub>), SO<sub>2</sub>, NO<sub>2</sub> and PM. WHO guidelines try to limit the lowest possible values for these pollutants. For example WHO limit values for PM2.5 is 10 micrograms per cubic meter is annual mean and 25 micrograms per cubic meter is 24-hour mean and limit value for PM10 is 20 micrograms/m<sup>3</sup> annual mean and 50 micrograms/m<sup>3</sup> 24-hour mean [14] 1. "The 2005 WHO Air quality guidelines" offer global guidance on thresholds and limits for key air pollutants that pose health risks. The Guidelines indicate that by reducing particulate matter (PM10) pollution from 70 to 20 micrograms per cubic metre, we can cut air pollution-related deaths by around 15 percent [14].

United States' Clean Air Act (CAA), first enacted in 1970 and with major revisions in 1990, is a federal law which is defined as "The Act that regulates air emissions from area, stationary, and mobile sources" [1]. CAA . EPA is the administrator of CAA [12]. As required by law, EPA regulates emissions standards for vehicles, industries, aircrafts and powerplants among others in order to protect environment and public health. Today, with the availability of new technology and analytical tools air quality data from the monitors across the regions can be accessed in an instant and can be analyzed for daily reporting. Based on daily AQI value, respective authorities can take appropriate actions to save outdoor air quality in areas where pollution level is insignificant and to identify measures to be taken in areas where air quality is poor.

## 8 METHODS

### 8.1 Air Quality Dataset

Outdoor air quality data sets are available in the USEPA.gov website called 'Air Data'. The data on 'Air Data website comes from AQS database where outdoor air data generated from thousands of air quality monitors from all over the country is collected. As mentioned, all states, local and private monitoring agencies send outdoor pollutants concentrations measurement data to the AQS database [1]. Besides, Air Data there are other sources of data as well, they are briefly discussed below;

- 'Air Now' which has air quality forecasts and real-time data in visual form.
- 'AirCompare' that has data about Counties' AQI summaries.
- 'AirTrends' data is about trends of air quality and emissions.
- 'Air Emissions Sources' has emissions data with national, state and county-level summaries for criteria pollutant emissions.
- 'Remote Sensing Information Gateway (RSIG)' that has air quality monitoring, monitoring and satellite data.
- 'Air Data' datasets have raw dataset, AQI summary datasets. Summary reports consists of AQI report which displays a

- yearly summary of AQI values in a county or city or Core Based Statistical Area (CBSA). The AQI values are summarized by maximum percentile and median and count of days in each AQI category and the count of days when AQI could be attributed to each criteria pollutant [1].
- “ Quality Statistics Report has yearly summaries of air pollution values for a city or county. It shows the maximum values reported during the year by all monitors in CBSA or county” [1].
  - Monitor Values Report that has yearly summary of the measurements at individual monitors and has descriptive information about the site [1].
  - Monitor Values Report-Hazardous Air Pollutants that shows HAPs summary data for individual monitoring sites [1].
  - Air Quality Index Daily Values Report that has information about AQI values for specified year and location [1].

The dataset that are being used for analysis are “Daily AQI by CBSA 2016” and “daily AQI by CBSA 2006”, from EPA’s air data website. Air data has different categories of outdoor air quality data. There are datasets for hourly as well as monthly time period broken down by single criteria pollutant of ‘Hazardous Air Pollutants’ (HAPs). There are county level monitors and stations and datasets grouped by “Core Based Statistical Area (CBSA)”. Each datasets have more than 17,000 data points.

CBSA is designed by Office of Management (OMB) as a geographical area that consists of one or more than one counties and similar surroundings that are associated with at least one core urbanized area of at least 10,000 population plus adjacent counties which are associated with each other in terms of social, economic and daily commutes [5]. CBSA collectively refers to Metropolitan and Micropolitan statistical areas. “OMB defined Metropolitan and Micropolitan statistical areas in 2003 based on application of the 2000 standards with Census 2000 data. It became effective in 2003” [5]. There are 922 CBSAs in total [5]. Metropolitan statistical areas are urbanized areas with population of 50,000 and its adjacent areas while Micropolitan statistical areas are areas with population of at least 10000 or less than 50,000 [5].

CBSA AQI datasets fits the scenario for monitoring outdoor air quality. Because the designed statistical areas are significantly populated along with higher concentrations of motor vehicles running, higher number of day to day activities, less natural habitats or and most of the areas are within industrial areas and powerplant generators. Also, the first look at the dataset give a general information about AQI value of each ‘Criteria Pollutant’ for a day in a year of each monitoring stations in CBSAs.

## 8.2 Methods

The comma separated dataset “daily aqi by cbsa 2016” was downloaded from data source “<https://aqs.epa.gov/aqsweb/airdata>”. The dataset shows AQI value of each “criteria pollutants” for each CBSA recorded per day per station for the year 2016. Criteria pollutants recorded in the dataset are, PM2.5, PM10, Ozone, SO2, NO2 and CO. It also has a column for number of stations for each CBSA and location of the monitoring stations per CBSA.

In order to have a comparative study of any changes in AQI value for each parameter for the listed CBSA, dataset for the year

2006 is also being analyzed. This gives us a picture of changes if any for the duration of 10 years. Since real world datasets may not be perfect, there are slight or negligible amount of discrepancies among the two datasets. Criteria pollutants recorded in the dataset varies within each CBSAs depending on their significance in the region or monitoring stations. Similarly, record date per station per CBSA is not continuous, there are certain interval for recording and reporting data. For example, data is recorded on January 1st 2016 and the next date is January 3rd and 6th and so on. This pattern is seen in the whole dataset. Also, number of monitoring stations varies per CBSAs.

Using jupyter notebook with python2.7 as the interpreter, the dataset is fetched and converted into pandas dataframe for and analysis. Next, only the columns needed for analysis are selected and created a clean pandas dataframe ready for manipulation. Jupyter Notebook is an open source web application which is powerful in data cleaning, manipulation, data analysis and visualizations. The notebook is not sufficient in itself for a variety of data manipulations, it needs to have all sorts of python packages as per requirement of data analysis. Matplotlib, pandas, numpy and pandas datetime are the packages used for air quality data analysis.

## 9 ANALYSIS

The requirements for the analysis are python’s jupyter notebook and the packages pandas, matplotlib and numpy. Using the pandas dataframe, average AQI value is calculated for each ‘Defining Parameter’ grouped by CBSA. Since AQI value determines the level of risk factor as shown in figure 2, it is worth calculating the mean of that value which helps in determining which CBSA is affected by which ‘defining Parameter’. It also helps identify the source of the pollutant so that responsible stakeholders can take required actions to solve the problem. The purpose of using the data set is to find out level of AQI per CBSA for the year 2016. The reason of using 2016 data set is it the most recent complete set of data for a year. While 2017 data set would have been the most recent look at AQI level in the United States but as of this analysis the 2017 data set contains air quality data until the month of May 2017. This data set wold be completed as a full years data only in the upcoming spring [6]. In order to do a general comparison of the changes, positive or negative, if any similar data set for the year 2006 is selected. This would allow us to look at differences in AQI values in a decade time frame.

### 9.1 CBSA Average AQI for 2016

Simple statistical measures such as mean, maximum, count and minimum value for any variable provides some insights into the degree of variation between each measure of the variable within a period of time. So, as a first test mean value of “Criteria Pollutants” for all the listed CBSAs have been calculated. Then the results have been plotted into a bar-chart as shown in figure 5. It shows that among all, Ozone and PM 2.5 have highest average among CBSAs while CO, NO2, SO2 and PM10 have slightly lower average AQI for the year 2016.

Furthermore, figure 3 illustrates the average AQI value for overall CBSAs for the year 2016 by month. The table illustrates that AQI value, which is 53.56, for Ozone is the highest during the month of

June of last year. This AQI value of ozone comes under the category of 'unhealthy for sensitive groups'. This value is followed by CO which has the highest mean value for two consecutive months, May and June. From the table it can be said that during 2016 the most significant criteria pollutants were Ozone, CO, and PM2.5.

In order to look at monthly average AQI per 'Defining Parameter', first 'Date' series is converted into pandas datetime format and then the series is bucketed into month of the year column. Then mean is calculated by using 'groupby' function, mean is calculated grouped by 'Defining Parameter' and 'month of year'. The result illustrated in figure 7 shows that there is no significant change in mean AQI for SO2 and PM10 while significant change can be seen for CO and Ozone. AQI average for CO shows a sharp increase in May.

For the purpose of finding out CBSA with highest and lowest average AQI value for the year 2016, for specified pollutant, aggregate function is used with groupby function with descending value of AQI. The result is shown in figure 9. The table shows CBSA 'Madison, WI' has lowest AQI value for SO2, followed by Fayetteville, NC for the same parameter. Similarly, as seen figure 10 Hilo, Hawaii has the highest AQI value for parameter SO2 which is 151.79. This means AQI value is unhealthy and people may experience health effects with sensitive groups people with asthma, bronchitis and chronic obstructive pulmonary disease (COPD) might have very serious health effects as stated in figure 2. The reason for highest AQI value for SO2 might be the release of SO2 gas from active volcanoes around the CBSA region as SO2 is one of the gases released from volcanoes. Hilo, HI is followed by 'Riverside– San Bernardino–Ontario, CA with AQI value 126.53 for 2 with category 'unhealthy'. Other CBSAs in the top ten list are Lansing–East Lansing, MI, San Juan–Carolina–Caguas, PR, Bishop, CA, Durango, CO, Riverside–San Bernardino–Ontario, CA, Philadelphia–Camden–Wilmington, PA–NJ–DE–MD, Los Angeles–Long Beach–Anaheim, CA and Minneapolis–St.Paul–Bloomington, MN–WI. All of these CBSAs have at least one defining parameter with 'Unhealthy' AQI value. It is also significant that Riverside–San Bernardino–Ontario, CA have two highest AQI value for two defining parameters that is ozone and PM 10.

In order to have a visual picture of level of AQI value, count measure is used grouped by 'CBSA', 'Defining Parameter' and 'Category' and visualized the result in a boxplot. This basically counted number of days in each category of AQI. The figure 11 shows, that there are significant number of 'good days' during 2016 followed by 'moderate', 'unhealthy for sensitive groups' and rest of the three categories have less range of days. As seen in bar chart the mean value for Ozone is higher for the months May and August in comparison to other pollutants Columns for the months of June and July are missing at this point.

## 9.2 Comparative AQI for 2006 and 2016

To compare AQI per CBSA for criteria pollutants, exactly similar data set for the year 2006 is used to calculate average AQI value and plotted into a bar chart which is shown in figure 6. As the result can be compared with that of 2016. average AQI value for ozone and SO2 is significantly higher compared to 2016 while that of carbon monoxide (CO) has increased significantly from 2006 to 2016. For other parameters there are not so significant changes.

Similarly, average AQI for SO2 is also higher in 2006 compared to 2016. There is also some changes in overall average AQI value. This figure shown here 4 illustrates the AQI average for the year 2006. As seen in barplot this figure shows the highest mean for Ozone at the top of the list when mean is grouped by month of the year and output is sorted in descending order by mean. Interestingly, most of the highest mean value are for the month of June, July and August. To be more specific AQI value for Ozone is mostly during above listed months. Similar result can also be seen in the analysis of 2016 dataset.

While comparing top ten CBSA with highest average AQI value for 2016 with that of 2006, top ten CBSAs with highest average AQI for 2006 is shown in figure 8. Compared to 2016 there is no sharp increase in mean value for CO any of the month as has been in 2016. Its mean value remains comparatively in same range, in between 17 and 16. Whereas, the value for Ozone is significantly higher for the months of March, April, May, August and September. For other months as well it has higher level of mean AQI compared to other pollutants except for SO2 which has the second highest AQI throughout the months. Average value for all of the pollutants throughout the months have higher values compared to that of year 2016 except for CO. Average CO AQI for 2016 is higher than that of the year 2006.

Another comparison can be done by looking at the list of top most CBSAs with higher mean AQI. We have already discussed about this for year of 2016. Now we will look at similar output for 2006 which is shown in figure 12. The mean output is generated by grouping the CBSA and defining parameter. Here, Bishop, CA has highest yearly average for PM10 which is 435.21 followed by CBSA Phoenix–Mesa–Scottsdale, AZ, Carlsbad–Artesia, NM have yearly mean value of 127 for SO2. In comparison to the same statistics for year 2016, mean value is lower than that of 2006. This indicated that there is progress made in lowering the measurement in air pollutant.

## 10 FURTHER STUDIES

It would be effective to look at the factors that helped lower the mean AQI value within a decade. Are there any policy differences between now and then? Are there less economic activities compared to 2016 or did any of the technology have any role to play to make such change? It would be effective to look at the factors that helped lower the mean AQI value within a decade. As can be seen from the analysis above average ozone has highest average value among all. Finding out or looking at the contributing factors can be another area of furthering this research. and it would be interesting to see any changes if any during the year of 2017. The analysis presented here showcases just the level of pollutants by AQI value. For furthering the studies different models for lowering AQI value for a single pollutant within certain periods of time can be developed. Since vehicular, industrial or air transportation emissions are some of the main sources of criteria pollutants, it would be very effective to analyze monthly emissions data with AQI data to have an understanding of level of pollutants generated from these sources. This analysis can be combined with weather data in order to find out the effects of weather or seasonal variations on increasing or lowering pollutants. There is an increased AQI value

for CO from 2006 to 2016. Factors responsible for poor AQI value can be the next research topics.

Looking at difference in emissions policy during the past decade will also provide some insight into why certain parameters have lower AQI value. Looking at effects of AQI value in international boundaries that is effect evaluation of pollutants of country affecting the air quality of surrounding country could be a whole new topic of outdoor air quality study.

## 11 CONCLUSION

Air quality monitors across the United States and its territories, Puerto Rico and Virgin Islands collects everyday air data in order to collect data on pollutants present in outdoor air. These generates a huge volume of data. Air data contains measurements of concentrations of pollutants, commonly called as criteria pollutants, present in outdoor air. These monitors are under the management of state, local or private environmental agencies. They collectively send these data to a national database system called air quality system. For analytical purpose the air data from the AQS website which is an source of air data, is directly accessed through jupyter notebook. Simple statistical measures are calculated to have a look at the level of pollutants present in ambient year.

The air data by CBSA, for the year of 2016 have been analyzed to see the average value of AQI for each defining parameter or criteria pollutants. This analysis identifies that average AQI value for Ozone is significantly higher compared to other pollutants. The analysis also shows that AQI value varies by month of the year. The results shows that Ozone AQI is higher in the months of March to August. While calculating overall average AQI, among all of the criteria pollutants level of ground level ozone has the highest value during the year of 2016 as well as 2006. It can also be concluded that overall level of AQI value have been improved over the years period while comparing the value from 2006 with that of 2016. Interesting enough AQI value of CO have been significantly increased from 2006 to 2016.

Air quality monitoring have become an important and effective policy for reducing the concentrations of different criteria pollutants. Since the adoption of clean air act in United States significant progress has been made in improving outdoor air quality. Thereby reducing the development of health risk factors for general public. Air quality monitoring stations collect everyday measurement values of criteria pollutants which generates a big volume of air data. This gives us information on daily level of pollutants present in the air thereby letting us know the quality of air that we breathe everyday. This information is shared to the public through various media outlet. This raise awareness among general public about the importance of outdoor air and their health.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his motivating words and attitude and support to achieve our best. Also I would like to express my gratitude to teaching assistants Juliette Zerick, Saber and Miao for their encouraging attitude and technical support and suggestions while writing this paper.

## REFERENCES

- [1] United States Environmental Protection Agency. 2017. Overview of the Clean Air Act and Air Pollution. Webpage. (April 2017). <https://www.epa.gov/clean-air-act-overview>
- [2] Mike Hamborg. 2017. *U.S. EPA and OpenAQ air quality data now available in BigQuery*. Technical Report. Google Cloud Platform. <https://cloud.google.com/blog/big-data/2017/06/us-epa-and-openaq-air-quality-data-now-available-in-bigquery>
- [3] Alexander Howard. 2015. *How IBM Is Using Big Data To Battle Air Pollution In Cities*. Report. HuffPost. [https://www.huffingtonpost.com/entry/ibm-big-data-air-pollution\\_us\\_56684e44e4b080edd565510](https://www.huffingtonpost.com/entry/ibm-big-data-air-pollution_us_56684e44e4b080edd565510)
- [4] Lucas Laursen. 2016. *AI and Big Data vs. Air Pollution*. Technical Report. IEEE-org. <https://spectrum.ieee.org/energy/environment/ai-and-big-data-vs-air-pollution>
- [5] United States Census Bureau. 2012. Geography. webpage. (December 2012). [https://www.census.gov/geo/reference/gtc/gtc\\_cbsa.html](https://www.census.gov/geo/reference/gtc/gtc_cbsa.html)
- [6] United States Environmental Agency. 2017. Air Data Basic Information. webpage. (October 2017). <https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information#what>
- [7] United States Environmental Protection Agency. 2016. webpage. (August 2016). <https://airnow.gov/index.cfm?action=aqibasics.aqi>
- [8] United States Environmental Protection Agency. 2016. Air Quality Management Process. webpage. (August 2016). <https://www.epa.gov/air-quality-management-process/managing-air-quality-ambient-air-monitoring>
- [9] United States Environmental Protection Agency. 2017. Health Effects of Ozone Pollution. webpage. (January 2017). <https://www.epa.gov/ozone-pollution/health-effects-ozone-pollution>
- [10] United States Environmental Protection Agency. 2017. Health Effects of Ozone Pollution. webpage. (January 2017). <https://www.epa.gov/so2-pollution/sulfur-dioxide-basics#effects>
- [11] WHO. 2005. Air quality guidelines global update 2005. Webpage. (2005). [http://www.who.int/phe/health\\_topics/outdoorair/outdoorair\\_aqg/en/](http://www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/)
- [12] Wikipedia. 2016. Clean Air Act (United States). webpage. (2016). [https://en.wikipedia.org/wiki/Clean\\_Air\\_Act\\_\(United\\_States\)](https://en.wikipedia.org/wiki/Clean_Air_Act_(United_States))
- [13] World Health Organization. 2005. *Air quality guidelines Global update 2005 Particulate matter and ozone and nitrogen dioxide and sulfur dioxide* (2005 ed.). WHO, UN City Marmorvej 51 DK-2100 Copenhagen Denmark. <http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/pre2009/air-quality-guidelines-global-update-2005>.
- [14] World Health Organization. 2016. Ambient (outdoor) air quality and health. Webpage. (September 2016). <http://www.who.int/mediacentre/factsheets/fs313/en/>

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

#### LIST OF FIGURES

1	WHO Guidelines Source And Health Effects [13]	8
2	Air Quality Index [7]	9
3	Top Ten Mean AQI Value for the Year 2016	10
4	Top Ten Mean AQI Value for the Year 2006	11
5	CBSA Average AQI for Criteria Pollutants	12
6	CBSA Average AQI for Criteria Pollutants 2006	13
7	Mean AQI per Month per Criteria Pollutant 2016	13
8	Mean AQI per Month per Criteria Pollutant 2006	14
9	Lowest Mean AQI per CBSA 2016	14
10	Highest Mean AQI per CBSA 2016	15
11	Boxplot Grouped by Category, 2016	16
12	Highest Mean AQI per CBSA 2006	17

Pollutant	Source types and major sources	Health effects	WHO guidelines
Particulate matter	Primary and secondary- Anthropogenic: burning of fossil fuel, wood burning, natural sources (e.g., pollen), conversion of precursors (NO <sub>x</sub> , SO <sub>x</sub> , VOCs) Biogenic: dust storms, forest fires, dirt roads	Respiratory symptoms, decline in lung function, exacerbation of respiratory and cardiovascular disease (e.g., asthma), mortality	PM <sub>10</sub> Annual mean: 20 µg/m <sup>3</sup> 24-hour mean: 50 µg/m <sup>3</sup> PM <sub>2.5</sub> Annual mean: 10 µg/m <sup>3</sup> 24-hour mean: 25 µg/m <sup>3</sup>
Ozone	Secondary- Formed through chemical reactions of anthropogenic and biogenic precursors (VOCs and NO <sub>x</sub> ) in the presence of sunlight	Decreased lung function, increased respiratory symptoms, eye irritation, bronchoconstriction	8-hour mean: 100 µg/m <sup>3</sup>
Nitrogen dioxide	Primary and secondary- Anthropogenic: fossil fuel combustion (vehicles, electric utilities, industry), kerosene heaters Biogenic: biological processes in soil, lightning	Decreased lung function, increased respiratory infection Precursor to ozone. Contributes to PM and acid precipitation	Annual mean: 40 µg/m <sup>3</sup> 1-hour mean: 200 µg/m <sup>3</sup>
Sulfur dioxide	Primary Anthropogenic: combustion of fossil fuel (power plants), industrial boilers, household coal use, oil refineries Biogenic: decomposition of organic matter, sea spray, volcanic eruptions	Lung impairment, respiratory symptoms. Precursor to PM. Contributes to acid precipitation	Annual mean: 20 µg/m <sup>3</sup> ; 10-minute mean: 500 µg/m <sup>3</sup>

Figure 1: WHO Guidelines Source And Health Effects [13]

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Figure 2: Air Quality Index [7]

<b>month_year</b>	<b>Defining Parameter</b>	<b>Mean</b>
2016-06	Ozone	53.56
2016-05	CO	51.96
2016-06	CO	51.61
2016-05	Ozone	49.07
2016-07	Ozone	48.35
2016-04	Ozone	47.86
2016-07	CO	47.27
2016-08	Ozone	44.83
2016-11	PM2.5	43.21
2016-01	PM2.5	42.18

Figure 3: Top Ten Mean AQI Value for the Year 2016

<b>month_year</b>	<b>Defining Parameter</b>	<b>Mean</b>
<b>2006-07</b>	<b>Ozone</b>	<b>69.58</b>
<b>2006-06</b>	<b>Ozone</b>	<b>69.29</b>
<b>2006-08</b>	<b>Ozone</b>	<b>61.31</b>
<b>2006-05</b>	<b>Ozone</b>	<b>58.31</b>
<b>2006-04</b>	<b>SO2</b>	<b>54.97</b>
	<b>Ozone</b>	<b>54.70</b>
<b>2006-08</b>	<b>PM2.5</b>	<b>51.20</b>
<b>2006-07</b>	<b>PM2.5</b>	<b>50.58</b>
<b>2006-05</b>	<b>SO2</b>	<b>50.36</b>
<b>2006-12</b>	<b>PM2.5</b>	<b>48.96</b>

Figure 4: Top Ten Mean AQI Value for the Year 2006

### CBSA Average AQI for Criteria Pollutants

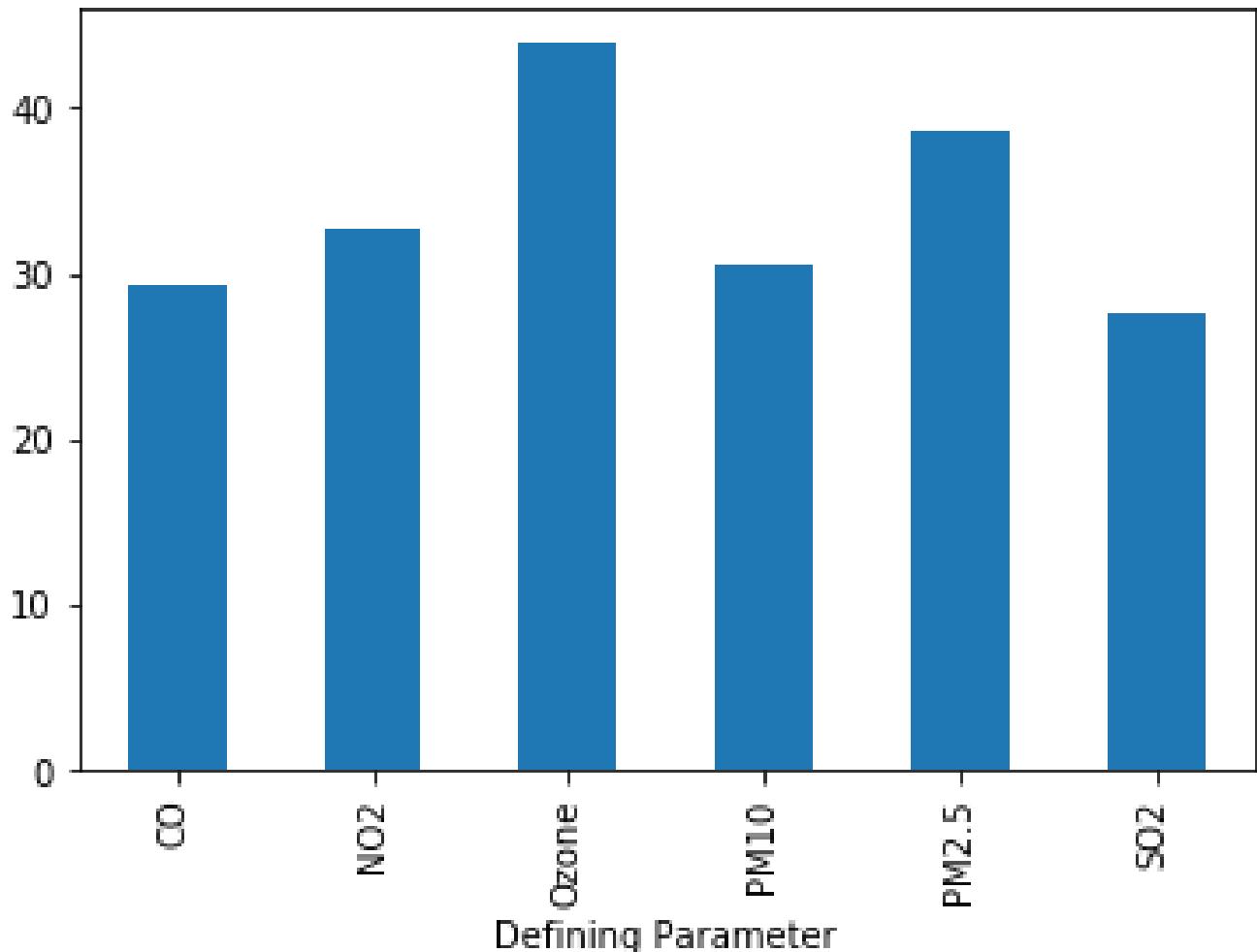


Figure 5: CBSA Average AQI for Criteria Pollutants

### CBSA Average AQI for Criteria Pollutants

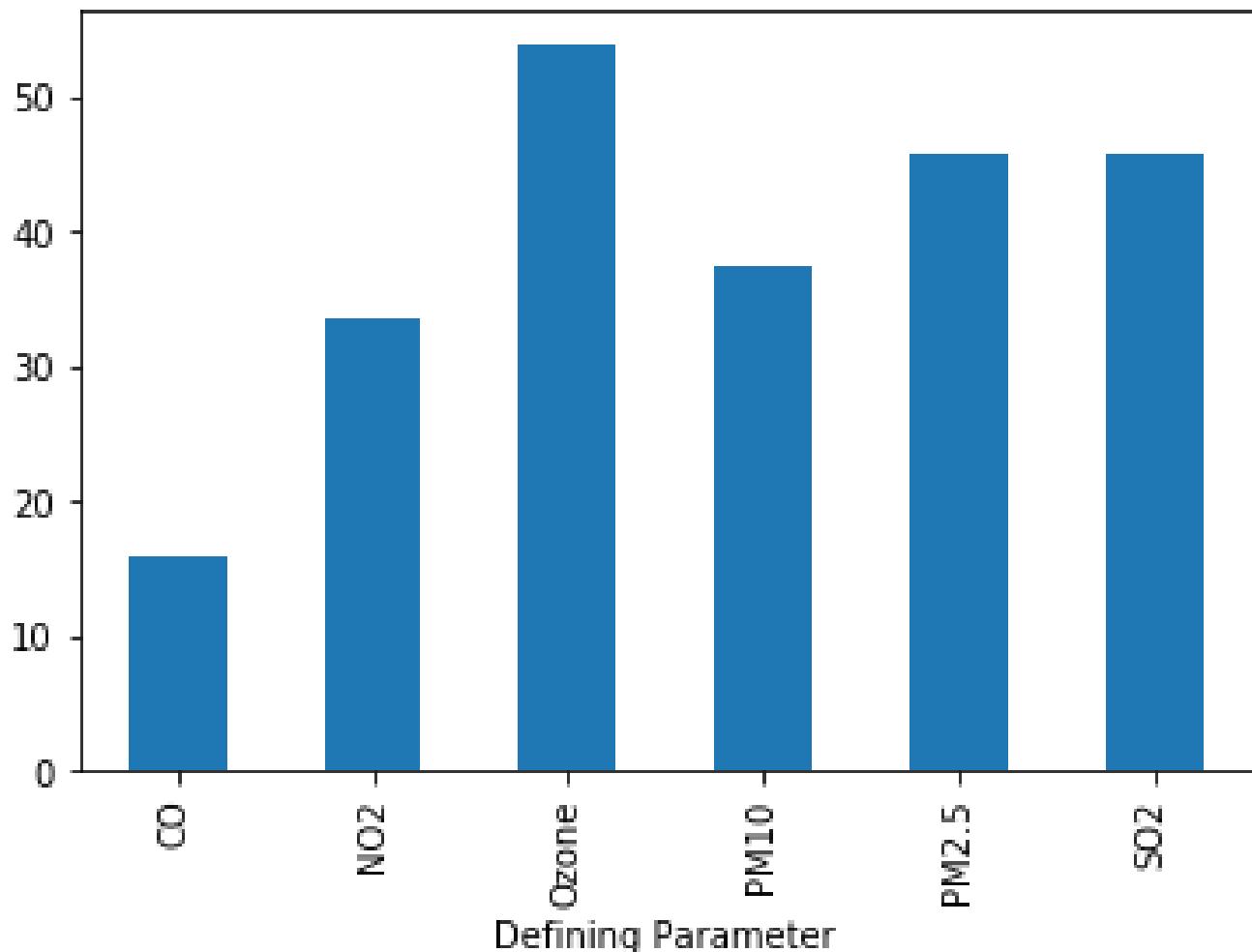


Figure 6: CBSA Average AQI for Criteria Pollutants 2006

Defining Parameters	2016-01	2016-02	2016-03	2016-04	2016-05	2016-08	2016-09	2016-10	2016-11	2016-12
CO	7.19	4.00	8.20	26.00	51.96	16.77	31.25	NA	26.00	2.60
NO <sub>2</sub>	33.18	33.07	33.69	39.00	34.41	22.90	40.43	31.27	33.11	30.39
Ozone	33.97	38.48	41.20	47.86	49.07	44.83	40.91	37.61	34.75	31.07
PM10	19.67	28.82	33.31	30.04	28.67	28.01	32.62	36.89	32.55	25.61
PM2.5	42.18	38.99	35.76	33.30	35.21	36.74	35.90	36.71	43.21	40.82
SO <sub>2</sub>	25.89	26.83	26.83	29.64	24.92	27.51	32.42	32.20	25.99	23.33

Figure 7: Mean AQI per Month per Criteria Pollutant 2016

Defining Parameters	2006-01	2006-02	2006-03	2006-04	2006-05	2006-08	2006-09	2006-10	2006-11	2006-12
CO	17.06	15.07	12.40	15.59	15.15	13.75	20.15	17.97	17.22	16.74
NO2	31.93	38.34	37.99	39.54	33.84	24.79	34.36	32.77	32.60	31.97
Ozone	32.99	37.80	46.22	54.70	58.31	61.31	46.37	38.06	33.51	29.79
PM10	45.46	48.20	32.75	36.07	36.64	31.49	37.91	36.17	40.79	33.55
PM2.5	43.23	45.92	43.81	39.97	43.43	51.20	47.00	41.82	47.47	48.96
SO2	43.84	44.25	44.65	54.97	50.36	48.75	45.89	47.61	43.86	42.40

Figure 8: Mean AQI per Month per Criteria Pollutant 2006

CBSA	Defining Parameter	mean
College Station-Bryan, TX	SO2	1.75
Wilmington, NC	SO2	1.44
Rochester, MN	SO2	1.00
Corning, NY	SO2	1.00
Jamestown-Dunkirk-Fredonia, NY	SO2	1.00
Kapaa, HI	SO2	0.50
Gulfport-Biloxi-Pascagoula, MS	SO2	0.50
Seneca, SC	SO2	0.46
Fayetteville, NC	SO2	0.38
Madison, WI	SO2	0.00

Figure 9: Lowest Mean AQI per CBSA 2016

CBSA	Defining Parameter	mean
Hilo, HI	SO2	151.79
Riverside-San Bernardino-Ontario, CA	Ozone	126.53
Lansing-East Lansing, MI	SO2	123.00
San Juan-Carolina-Caguas, PR	SO2	119.61
Bishop, CA	PM10	117.66
Durango, CO	NO2	109.00
Riverside-San Bernardino-Ontario, CA	PM10	108.20
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	SO2	107.00
Los Angeles-Long Beach-Anaheim, CA	Ozone	105.46
Minneapolis-St. Paul-Bloomington, MN-WI	SO2	105.00

Figure 10: Highest Mean AQI per CBSA 2016

Boxplot grouped by Category

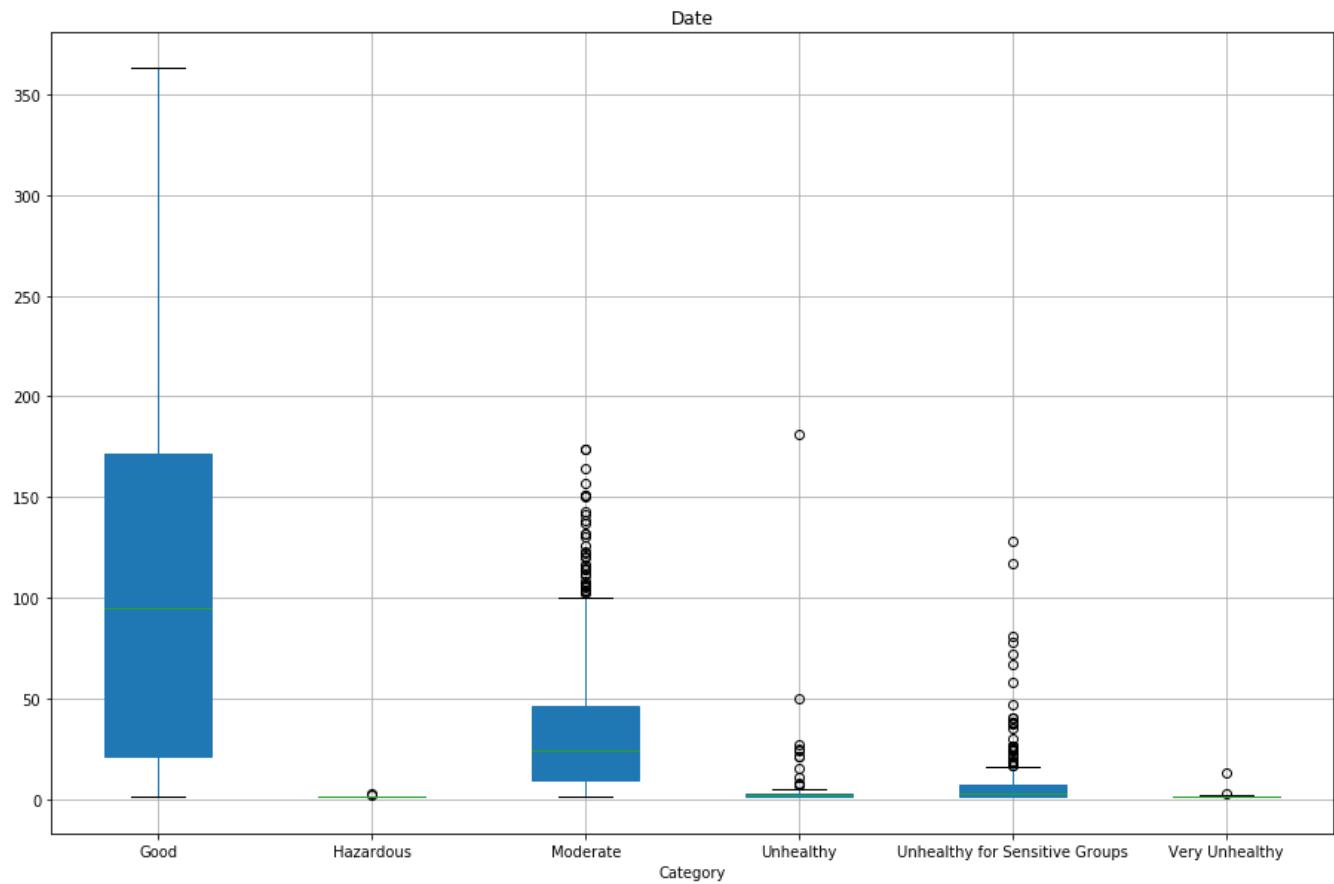


Figure 11: Boxplot Grouped by Category, 2016

CBSA	Defining Parameter	mean
Bishop, CA	PM10	435.21
Phoenix-Mesa-Scottsdale, AZ	PM10	185.33
Carlsbad-Artesia, NM	SO2	161.00
El Centro, CA	SO2	127.50
Riverside-San Bernardino-Ontario, CA	Ozone	123.17
Bakersfield, CA	Ozone	119.92
Atlanta-Sandy Springs-Roswell, GA	Ozone	119.33
Houston-The Woodlands-Sugar Land, TX	Ozone	118.44
Birmingham-Hoover, AL	Ozone	114.16
St. Louis, MO-IL	SO2	112.85

Figure 12: Highest Mean AQI per CBSA 2006

```
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
```

```
[2017-12-10 13.52.11] 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 1.3s.
./README.yml
9:81     error    line too long (86 > 80 characters) (line-length)
10:81    error    line too long (88 > 80 characters) (line-length)
24:81    error    line too long (92 > 80 characters) (line-length)
25:81    error    line too long (99 > 80 characters) (line-length)
26:81    error    line too long (98 > 80 characters) (line-length)
27:81    error    line too long (102 > 80 characters) (line-length)
28:81    error    line too long (100 > 80 characters) (line-length)
29:81    error    line too long (87 > 80 characters) (line-length)
30:81    error    line too long (103 > 80 characters) (line-length)
31:81    error    line too long (99 > 80 characters) (line-length)
32:81    error    line too long (99 > 80 characters) (line-length)
33:81    error    line too long (87 > 80 characters) (line-length)
34:12    error    trailing spaces (trailing-spaces)
37:9     error    trailing spaces (trailing-spaces)
38:8     error    wrong indentation: expected 8 but found 7 (indentation)
46:81    error    line too long (103 > 80 characters) (line-length)
```

```
47:81    error    line too long (102 > 80 characters) (line-length)
48:81    error    line too long (103 > 80 characters) (line-length)
49:81    error    line too long (102 > 80 characters) (line-length)
50:81    error    line too long (101 > 80 characters) (line-length)
51:81    error    line too long (98 > 80 characters) (line-length)
52:81    error    line too long (98 > 80 characters) (line-length)
53:81    error    line too long (104 > 80 characters) (line-length)
54:81    error    line too long (100 > 80 characters) (line-length)
55:81    error    line too long (101 > 80 characters) (line-length)
56:81    error    line too long (103 > 80 characters) (line-length)
57:81    error    line too long (104 > 80 characters) (line-length)
58:81    error    line too long (102 > 80 characters) (line-length)
59:81    error    line too long (100 > 80 characters) (line-length)
60:81    error    line too long (106 > 80 characters) (line-length)
61:81    error    line too long (102 > 80 characters) (line-length)
62:81    error    line too long (99 > 80 characters) (line-length)
63:81    error    line too long (98 > 80 characters) (line-length)
64:81    error    line too long (93 > 80 characters) (line-length)
65:81    error    line too long (104 > 80 characters) (line-length)
66:81    error    line too long (101 > 80 characters) (line-length)
70:25    error    trailing spaces (trailing-spaces)
76:79    error    trailing spaces (trailing-spaces)
79:1     error    trailing spaces (trailing-spaces)
```

---

## Compliance Report

---

name: Janaki Mudvari Khatiwada  
hid: 330  
paper1: 100% Review date nov 1  
paper2: 100%  
project: 100%

yamlcheck

---

wordcount

---

17

wc 330 project 17 6041 report.tex  
wc 330 project 17 6089 report.pdf

wc 330 project 17 466 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

53: 'Very Unhealthy' and 'Hazardous' \cite{airnow-gov}. The AQI value range from 0 to 500. The agency has assigned colors ('Green', 'Yellow', 'Orange', 'Red', 'Purple' and 'Maroon' respectively) to each of the air quality categories \cite{airnow-gov}. It is shown in figure \ref{AQI}.

68: EPA has prioritized six major air pollutants that are commonly found all over the US. They pose significant threat to public health and environment. They are called 'Criteria Pollutants' and they are ground level ozone, fine particles or particulate matter (PM2.5 and PM10), nitrogen dioxide, sulphur dioxide, lead and carbon monoxide \cite{epa-gov}. WHO has set the guidelines for each of these pollutants as shown in figure\ref{WHOGuidelines}.

80: AQI is the index, for five major air pollutants discussed above, calculated by special formula developed by EPA. EPA uses its own formula to convert daily concentrations of measurement of each pollutant into AQI value of each pollutant \cite{airnow-gov}. Among all the highest AQI value is reported as the daily AQI value for that day \cite{airnow-gov}. Generally AQI 100 is the acceptable index set by EPA to protect public's health and it

ranges from 0 to 500 \cite{airnow-gov}. The higher the value of AQI, greater is the pollution level and greater is the health risk. Based on hourly data collection from air quality monitors, stakeholders can constantly monitor AQI value in their cities or respective location. So weather channels in different media outlets such as local radio, television stations and newspapers also report about AQI index in order to inform general public about air quality in their area. Figure \ref{AQI} shows the AQI classification for each pollutant as recommended by WHO and implemented by U. S. EPA. EPA is required to report any AQI value greater than 100 specifically in larger cities with population more than 350,000 \cite{airnow-gov}.

- 88: WHO guidelines for air quality is applied worldwide. This guidelines was revised in 2005 \cite{www-who}. The guidelines set standards for different air pollutants. According to the guidelines which is based on scientific evidence WHO has set standards for Ozone (O<sub>3</sub>), SO<sub>2</sub>, NO<sub>2</sub> and PM. WHO guidelines try to limit the lowest possible values for these pollutants. For example WHO limit values for PM2.5 is 10 micrograms per cubic meter is annual mean and 25 micrograms per cubic meter is 24-hour mean and limit value for PM10 is 20 micrograms/m<sup>3</sup> annual mean and 50 micrograms/m<sup>3</sup> 24-hour mean \cite{www-who} \ref{WHOGuidelines}. ‘‘The 2005 WHO Air quality guidelines’’ offer global guidance on thresholds and limits for key air pollutants that pose health risks. The Guidelines indicate that by reducing particulate matter (PM10) pollution from 70 to 20 micrograms per cubic metre, we can cut air pollution-related deaths by around 15 percent \cite{www-who}.
- 131: The requirements for the analysis are python’s jupyter notebook and the packages pandas, matplotlib and numpy. Using the pandas dataframe, average AQI value is calculated for each ‘Defining Parameter’ grouped by CBSA. Since AQI value determines the level of risk factor as shown in figure \ref{AQI}, it is worth calculating the mean of that value which helps in determining which CBSA is affected by which ‘defining Parameter’. It also helps identify the source of the pollutant so that responsible stakeholders can take required actions to solve the problem. The purpose of using the data set is to find out level of AQI per CBSA for the year 2016. The reason of using 2016 data set is it the most recent complete set of data for a year. While 2017 data set would have been the most recent look at AQI level in the United States but as of this analysis the 2017 data set contains air quality data until the month of May 2017. This data set would be completed as a full years data only in the upcoming spring \cite{outdoor-air}. In order to do a general comparison of the changes, positive or negative, if any similar data set for the

year 2006 is selected. This would allow us to look at differences in AQI values in a decade time frame.

- 133: Simple statistical measures such as mean, maximum, count and minimum value for any variable provides some insights into the degree of variation between each measure of the variable within a period of time. So, as a first test mean value of ‘‘Criteria Pollutants’’ for all the listed CBSAs have been calculated. Then the results have been plotted into a bar--chart as shown in figure \ref{Average CBSA AQI 2016}. It shows that among all, Ozone and PM 2.5 have highest average among CBSAs while CO, NO<sub>2</sub>, SO<sub>2</sub> and PM10 have slightly lower average AQI for the year 2016.
- 135: Furthermore, figure \ref{Top Ten Mean AQI Value for the Year 2016} illustrates the average AQI value for overall CBSAs for the year 2016 by month. The table illustrates that AQI value, which is 53.56, for Ozone is the highest during the month of June of last year. This AQI value of ozone comes under the category of ‘unhealthy for sensitive groups’. This value is followed by CO which has the highest mean value for two consecutive months, May and June. From the table it can be said that during 2016 the most significant criteria pollutants were Ozone, CO, and PM2.5.
- 137: In order to look at monthly average AQI per ‘Defining Parameter’, first ‘Date’ series is converted into pandas datetime format and then the series is bucketed into month of the year column. Then mean is calculated by using ‘groupby’ function, mean is calculated grouped by ‘Defining Parameter’ and ‘month of year’. The result illustrated in figure \ref{Mean AQI per Month 2016} shows that there is no significant change in mean AQI for SO<sub>2</sub> and PM10 while significant change can be seen for CO and Ozone. AQI average for CO shows a sharp increase in May.
- 139: For the purpose of finding out CBSA with highest and lowest average AQI value for the year 2016, for specified pollutant, aggregate function is used with groupby function with descending value of AQI. The result is shown in figure \ref{Lowest Mean AQI per CBSA 2016}. The table shows CBSA ‘Madison, WI’ has lowest AQI value for SO<sub>2</sub>, followed by Fayetteville, NC for the same parameter.
- 140: Similarly, as seen figure \ref{Highest Mean AQI per CBSA 2016} Hilo, Hawaii has the highest AQI value for parameter SO<sub>2</sub> which is 151.79. This means AQI value is unhealthy and people may experience health effects with sensitive groups people with asthma, bronchitis and chronic obstructive pulmonary disease (COPD) might have very serious health effects as stated in figure \ref{AQI}. The reason for highest AQI value for SO<sub>2</sub> might be the release of SO<sub>2</sub> gas from active volcanoes around the CBSA region as SO<sub>2</sub> is one of the gases released from volcanoes. Hilo, HI is followed by ‘Riverside-- San Bernardino--Ontario, CA with AQI

value 126.53 for \ref{AQI} with category ‘unhealthy’. Other CBSAs in the top ten list are Lansing--East Lansing, MI, San Juan --Carolina--Caguas, PR, Bishop, CA, Durango, CO, Riverside--San Bernardino--Ontario, CA, Philadelphia--Camden--Wilmington, PA--NJ --DE--MD, Los Angeles--Long Beach--Anaheim, CA and Minneapolis--St.Paul--Bloomington, MN--WI. All of these CBSAs have at least one defining parameter with ‘Unhealthy’ AQI value. It is also significant that Riverside--San Bernardino--Ontario, CA have two highest AQI value for two defining parameters that is ozone and PM 10.

- 142: In order to have a visual picture of level of AQI value, count measure is used grouped by ‘CBSA’, ‘Defining Parameter’ and ‘Category’ and visualized the result in a boxplot. This basically counted number of days in each category of AQI. The figure \ref{boxplot} shows, that there are significant number of ‘good days’ during 2016 followed by ‘moderate’, ‘unhealthy for sensitive groups’ and rest of the three categories have less range of days. As seen in bar chart the mean value for Ozone is higher for the months May and August in comparison to other pollutants Columns for the months of June and July are missing at this point.
- 147: To compare AQI per CBSA for criteria pollutants, exactly similar data set for the year 2006 is used to calculate average AQI value and plotted into a bar chart which is shown in figure \ref{Average CBSA AQI 2006}. As the result can be compared with that of 2016.
- 149: There is also some changes in overall average AQI vlaue. This figure shown here \ref{Top Ten Mean AQI Value for Year 2006} illustrates the AQI average for the year 2006. As seen in barplot this figure shows the highest mean for Ozone at the top of the list when mean is grouped by month of the year and output is sorted in descending order by mean. Interestingly, most of the highest mean value are for the month of June, July and August. To be more specific AQI value for Ozone is mostly during above listed months. Similar result can also be seen in the analysis of 2016 dataset.
- 151: While comparing top ten CBSA with highest average AQI value for 2016 with that of 2006, top ten CBSAs with highest average AQI for 2006 is shown in figure \ref{Mean AQI per Month 2006}. Compared to 2016 there is no sharp increase in mean value for CO any of the month as has been in 2016. Its mean value remains comparatively in same range, in between 17 and 16. Whereas, the value for Ozone is significantly higher for teh months of March, April, May, August and September. For other months as well it has higher level o mean AQI compared to other pollutants except for SO<sub>2</sub> which has the second highest AQI throughout the months.

Average value for all of the pollutants throughout the months have higher values compared to that of year 2016 except for CO.

Average CO AQI for 2016 is higher than that of the year 2006.

- 153: Another comparison can be done by looking at the list of top most CBSAs with higher mean AQI. We have already discussed about this for year of 2016. Now we will look at similar output for 2006 which is shown in figure \ref{Highest Mean AQI per CBSA 2006}. The mean output is generated by grouping the CBSA and defining parameter. Here, Bishop, CA has highest yearly average for PM10 which is 435.21 followed by CBSA Phoenix-Mesa-Scottsdale, AZ, Carlsbad-Artesia, NM have yearly mean value of 127 for S02. In comparison to the same statistics for year 2016, mean value is lower than that of 2006. This indicated that there is progress made in lowering the measurement in air pollutant.
- 187: \begin{figure}[htb]
- 188: \includegraphics[width=1.0\columnwidth]{images/sourceandhealtheffects.png}
- 190: \label{WHOGuidelines}
- 193: \begin{figure}[htb]
- 194: \includegraphics[width=1.0\columnwidth]{images/aqiclassification.png}
- 196: \label{AQI}
- 200: \begin{figure}[htb]
- 201: \includegraphics[width=1.0\columnwidth]{images/top10meanaqi2016.png}
- 203: \label{Top Ten Mean AQI Value for the Year 2016}
- 206: \begin{figure}[htb]
- 207: \includegraphics[width=1.0\columnwidth]{images/top10meanaqi2006.png}
- 209: \label{Top Ten Mean AQI Value for Year 2006}
- 213: \begin{figure}[htb]
- 214: \includegraphics[width=1.0\columnwidth]{images/averageaqi2016.png}
- 216: \label{Average CBSA AQI 2016}
- 219: \begin{figure}[htb]
- 220: \includegraphics[width=1.0\columnwidth]{images/averageaqi2006.png}
- 222: \label{Average CBSA AQI 2006}
- 225: \begin{figure}[htb]
- 226: \includegraphics[width=1.0\columnwidth]{images/avaqibymonth2016.png}
- 228: \label{Mean AQI per Month 2016}
- 232: \begin{figure}[htb]
- 233: \includegraphics[width=1.0\columnwidth]{images/avaqibymonth2006.png}
- 235: \label{Mean AQI per Month 2006}

```
239: \begin{figure}[htb]
240: \includegraphics[width=1.0\columnwidth]{images/lowestmeanaqipercent
sa.png}
242: \label{Lowest Mean AQI per CBSA 2016}
245: \begin{figure}[htb]
246: \includegraphics[width=1.0\columnwidth]{images/cbsahighestaqi.png
}
248: \label{Highest Mean AQI per CBSA 2016}
251: \begin{figure}[htb]
252: \includegraphics[width=1.0\columnwidth]{images/boxplotofcountofaq
ibycategory2016.png}
254: \label{boxplot}
257: \begin{figure}[htb]
258: \includegraphics[width=1.0\columnwidth]{images/top10cbsamean2006.
png}
260: \label{Highest Mean AQI per CBSA 2006}
```

```
figures 12
tables 0
includegraphics 12
labels 12
refs 15
floats 12
```

```
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
```

---

WARNING: table and above may be used improperly

80: AQI is the index, for five major air pollutants discussed above, calculated by special formula developed by EPA. EPA uses its own formula to convert daily concentrations of measurement of each pollutant into AQI value of each pollutant \cite{airnow-gov}. Among all the highest AQI value is reported as the daily AQI value for that day \cite{airnow-gov}. Generally AQI 100 is the acceptable index set by EPA to protect public's health and it ranges from 0 to 500 \cite{airnow-gov}. The higher the value of AQI, greater is the pollution level and greater is the health risk. Based on hourly data collection from air quality monitors, stakeholders can constantly monitor AQI value in their cities or respective location. So weather channels in different media outlets such as local radio, television stations and newspapers also report about AQI index in order to inform general public about air quality in their area. Figure \ref{AQI} shows the AQI classification for each pollutant as recommended by WHO and implemented by U. S. EPA. EPA is required to report any AQI value greater than 100 specifically in larger cities with population more than 350,000 \cite{airnow-gov}.

WARNING: figure and above may be used improperly

149: There is also some changes in overall average AQI value. This figure shown here \ref{Top Ten Mean AQI Value for Year 2006} illustrates the AQI average for the year 2006. As seen in barplot this figure shows the highest mean for Ozone at the top of the list when mean is grouped by month of the year and output is sorted in descending order by mean. Interestingly, most of the highest mean value are for the month of June, July and August. To be more specific AQI value for Ozone is mostly during above listed months. Similar result can also be seen in the analysis of 2016 dataset.

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

# Agricultural Data Science: Then, Now, and Beyond

Ross Wood

HID345

rmw@indiana.edu

## ABSTRACT

As the human population swells to staggering numbers that historians of yesteryear could not imagine, one very important question seems to keep coming up over and over again. How do we feed all of these people? Thankfully, humans are intelligent beasts and are figuring out ways to farm and produce larger amounts of food using methods and techniques more sophisticated than ones humanity has relied on in the past. The party is just getting started as farming meets the era of big data. As more and more data is generated from farming, techniques and processes become more sophisticated, cleaner, and more efficient. The kind of data being analyzed to improve agricultural endeavours comes in many forms, and can be statistical data like amount of food grown using how much land, actual data generated from using farm tools and other smart farming equipment, or any other kind of agricultural activity that can produce datafies actions and procedures. However, data science is helping in other ways, too, as scientists and engineers are taking advantage of all this newly available data and helping create new technology to improve food production and increase yields. With all this new information available, new farming endeavours are being undertaken. Farming within closed systems such as urban or vertical farming, practicing precision agricultural techniques, or even laboratories using genetics data on different plant strains to crossbreed the various plant strains in order to produce new breeds that can grow in the harshest of environments while using minimal resources. As the population grows, we are finding that not only is the production of food vital, but also that sustainable farming techniques are of paramount importance for long term agricultural need. Data Science and its applications are most definitely changing the way people produce food and the very nature of farming itself.

## KEYWORDS

i523, HID345, Agricultural Data Science, Smart Farming, Vertical Farms, Urban Farming, Big Data Farming, Smart Farming Tools, Precision Agriculture

## 1 INTRODUCTION

Humans have not always lived in the amazing concrete and technological jungles that we have surrounded ourselves with today. Indeed, the ability to stop being nomadic and settle down in one area is a relatively new development in regards to the grand scale of human existence. However, if there is one technological advancement which is considered to be the most directly responsible for allowing humans to change how they lived and thrive in a harsh and unforgiving world, it would be when ancient humans evolved into the first agrarian societies by figuring out how to plant crops and grow food. By making their societies agriculture based instead of hunting based, humans were able to live in one place and do a lot of gathering, in addition to the hunting they were used to. This

regular supply of food and less dependency on hunting allowed ancient humans time to develop other aspects of human society, such as language, writing, and building. This was about 12,000 years ago and ever since that time, humanity has been gathering data on farming and slowly but surely refining the techniques we use for food production. Humanity has not just been gathering data and knowledge on how to grow food, but also information on what kinds of crops to grow when and where, how to deal with insect, rodent, and pests and other external threats to crops, and how what to do and how to manage different weather and environmental setbacks. These are a few of the many examples of information that humanity has accumulated over the millennia that have allowed humans to improve their farming and agricultural techniques, which has enabled humanity to thrive around the world.

The advances humans have made in their early years of farming will pale in comparison to the advances that humanity has the potential to make in the modern era using big data analytics and sophisticated technology that improves farming methods. With population statistics indicating that there is no sign of our population growth slowing down, the question is becoming even more relevant today than it has been in the past. With estimates putting the world population at 11.2 billion by the year 2100, solving the hunger problem is imperative [13]. Using data science and modern data analysis techniques and models, humans are able to look at the concept of agriculture as a whole and start making decisions based on data that had never been readily available in the past. This data is useful to farmers and serves as kind of risk management system through which farmers can make informed decisions about changes they might want to implement on their farms. The rise of smart farming with a focus on sustainability, the ability to analyze information in ways not possible in the past, and the invention of tools like monitoring sensors and machines that datafie work are changing agriculture for the better. The desire for more efficiency, greater food production, and meeting the demands of a steadily growing population are the driving forces behind why this field continues to be researched and expanded. Closed farms that eliminate the need for pesticides and rodenticide use while also helping the environment by cleaning the air in urban areas are a couple of unique ways this field has branched off from traditional farming practices and techniques.

Modern technology and data science are making new agricultural endeavours possible in ways that previously could not have been attempted with any hope of such success modern farmers are finding. One new technique is called urban, or vertical farming. A vertical farm is a farm constructed in an urban area that goes up instead of out. Ideally built in a parking garage like structure, vertical farms use data analysis to make their crop yields extremely efficient. There are plans for their construction coming up in more and more cities over the next couple decades and they are already beginning

to appear in developing nations that have problems meeting the demand for fresh fruits and vegetables in major urban centers. Other examples of how data science is changing the agriculture game is through the invention of more sophisticated computer software and languages that allow for the analysis of farming and crop data in ways that could never have been done in the past. This new way of looking at growing, storing, and transporting food and agricultural goods is making humanity second guess a lot of the ways we used to do things with regard to food. The explosive growth of big data and the rise of data science are already changing the way the world works and how we go about our daily lives. Data science is already improving agricultural endeavours in a myriad of ways, just like it does in most fields that it used to solve problems in. The positive benefits of this transition to new approaches in agriculture are already beginning to be seen.

## 2 HISTORICAL AGRICULTURAL DATA SCIENCE

The rise of big data analytics software and technology was the turning point where society began to really be able to take advantage of the ever growing amounts of data being produced by farms and their workers. As computer components began to get smaller, cheaper, and more powerful, this enabled more wide spread use of data analysis to be performed, which made it easier for companies and other organizations to adopt these new techniques and get in on the ground floor of solving agricultural problems and changing the way people farm around the world. Despite all these positive steps happening in the fields of technology, data analysis, and data creation, only small groups of people, usually limited to college university campuses, were actually receiving funding to analyze agricultural research data. This lack of funding meant that, although models for agricultural analysis were being developed, they were not being improved upon, leaving the power of data science in the field of agriculture unrealized until it could be adopted by more people and organizations [16].

### 2.1 Early Agricultural Analysis

In early to mid 1950s, computers and funding were still only available mostly at universities, which meant that universities were where most data analysis was being done in that era. Despite this, great steps were still being made to lay the foundation of modern agricultural data science and all that that encompasses. From the mid 1950s and onward, modeling was done by various institutions to find things like best water balance, photosynthesis and growth statistics, models to evaluate land and zoning properties, and pinning down economic risk management models, to name a few [16]. These models acted as guidelines and risk management tools in food production decisions making processes. As the availability and use of modeling technology became more widespread and more people and organizations began adopting them, so too did their positive impact on the real world become more apparent and visible. The changes, technological advances, and new techniques that were being proposed to farmers may have been met with some skepticism at first, but by the mid 1970s, these new processes were reportedly responsible for saving the lives of billions of people around the world by helping aid world hunger relief efforts. Like dominoes

falling, this led to increases in funding, which led to better and more accurate models being developed, which led to even more advances, which led to more funding, and so on. This self sustaining cycle of budget increases and technological innovation at the beginning of the agricultural data science boom led to the development of new ways of thinking about growing food and led to the development of tools that are still used today. These agricultural systems were developed for many reasons, but the top three reasons they are developed are typically “the intended use of the model, approaches taken to develop the models, and their target scales” [16]. Whatever their intent for development, very quickly it was evident that the proof was in the pudding.

### 2.2 New Tools for Agricultural and Risk Modeling

One data analysis tool for statistical decision making that was developed during the early days of agricultural data analysis was a software called Statistical Analysis System, or SAS for short. SAS was developed and “started out as a tool for statisticians: Goodnight originally developed it to analyze agricultural-research data in North Carolina” [16]. The development and utilization of SAS made sifting through the piles of random agricultural data much more manageable, and helped proto-data scientists find patterns, make connections, and gleam wisdom from the available agricultural data that they would not have been able to find without using SAS. Being able to manipulate and understand all the available data gave the farmers and data scientists a statistical edge when making choices about changing their farming practices. Economic risk management models would later be developed which helped make farmers more knowledgeable and able to hedge their bets whenever they were attempting new processes and techniques for improving their crops and growing capacity [25]. All of these new models made it easier for farmers to make decisions about making changes in how they produced their crops. The ability to make informed decisions about potentially big changes made farmers less reticent to try new approaches to farming and different techniques in how they produced their crops.

Since the development and implementation of SAS, other techniques have been developed to help take the pressure off of farmers when making educated and informed decisions. Commonly referred to as decision support tools, they have been developed to help farmers and data scientists make heads or tails of all the data that their trade is generating on a day to day basis. Thanks to the development of these kinds of tools and models and their improvement over the years, the decades since their adoption by farmers have seen farms produce more food per acre and increase their own sustainability for the years to come. These tools are incredibly useful to farmers, and “lead users through clear steps and suggest optimal decision paths or may act as information sources to improve the evidence base for decisions” [25]. These tools were slow to be developed because of funding problems, but once they began to catch on, they were quickly adopted by farmers to help improve their yields. A statistical analysis of farmers and the tools they used found that the odds of using decision support tools and software increased greatly depending on the size of the farm. The bigger the farm, the better the chances are that they use some kind of

decision support tools software [25]. The growth of urbanization and major cities meant that there were less farmers, which means that the farms that did exist were becoming bigger and bigger to fill the vacuum. Larger farms led to more widespread of adoption of the then modern technology and approaches to increase yields, improve efficiency, and improve farming conditions and overall food production.

### 2.3 Food Explosion, Courtesy of Data Science

Slowly but surely, the world, and especially developing nations, began to see the effects of applied data science to agricultural and farming endeavours. It was in the 1960s that a kind of critical mass was reached where the benefits of this field became undeniable. This led to the eventual development of precision farming with a focus on sustainability. All of these new breakthroughs in environmental science helped plant the seeds for what would be the growing environmental movements. In the mid to late 20th century, scientists like Norman Borlaug, who would later win a Nobel Peace Prize for his research, pioneered the way in using analysis models on plant genetics in order to improve yield by finding which breeds were best to cross and grow in different environments. These experiments led directly to the development of high yield crops in 1960s Mexico, and would later do the same in India [5]. Analysis of crop data allowed for the scientist to breed genetically superior strains of cereal grains that could withstand harsher climates, thrive on fewer resources like water and fertilizer, and produce a greater yield per acre than strains that Mexican farmers had previously been using. These new strains were potent and unlike anything the world had ever seen. When the Mexican farmers adopted these strains that Borlaug and his team had developed, they immediately began to see an improvement in their ability to meet the food demand of their population. While not alleviating the problem of hunger entirely, these developments proved that applying data science solutions to agricultural problems yields great results. These developments helped stymie off a hunger epidemic, and when Borlaug received his Nobel Peace Prize in 1970, it was officially for “saving over a billion lives” [4]. This quick turnaround showed definitively that research and development in agricultural data science was worth the investment, as less than 25 years after these new processes and models were being worked out, they were able to be used to save billions of lives.

These achievements, fantastical as they may be, only helped stymie off the threat of hunger around the world temporally, as these kinds of advances could only go so far. Growing populations steadily defeat advancements in food production, and humanity has to keep adapting and refining our methods and techniques in order to continually meet the needs of the many [22]. Luckily, humans are clever creatures, and our innovation and technological achievements have grown exponentially with our massive populations. More powerful computers and data analysis methods are constantly making humans better at producing food and meeting the staggering population demands. The gradual advance in sophistication of humankind’s ability to produce food and knowledge of best techniques and practices continues to evolve with technology and available data. Data science is just the newest and sharpest tool

in humanity’s toolbox and its use is improving farming in every way.

### 2.4 The End of the Past

History is a gradual process which only seems instantaneous when reading about it in a history book. When talking about agricultural data science and its past, present, and future, it would be easier to think about it like stepping stones. The future did not just come to be, but contributions from the farmers and data scientists in the past helped to build it up to where it is today. But which stepping stones are the most important ones to be looking at when thinking agricultural data science’s past? Ultimately, the best example the past can provide is that working together and openly is the most beneficial approach for everyone involved.

Some of the biggest developments from the 20th century of agriculture came about because of the open nature of the field, but there were other circumstances that led to great leaps in agricultural data science. Other circumstances that have led to advances in this field are: 1) the ability to capitalize on a crisis, 2) advances in technology and hardware, 3) keeping the data open and harmonized, 4) making the data easily applicable so that it can be used in other disciplines, 5) developing and maintaining standards and protocols, and 6) making sure the data remains user friendly and user-driven [16]. These approaches represent ideal ways to handle and manage agricultural data so that it can be used and expanded upon by all parties that might be interested. Keeping the data open allows for greater innovation as it becomes a problem that is solved on a societal level, not an individual one. This helps keep the data user friendly and driven as well. All of these different examples represent the various stepping stones that have worked together to bring the field of agricultural data science from its roots in human history and changes in the 20th century, to the modern way we look at agriculture and farming in the 21st century.

## 3 MODERN AGRICULTURAL DATA SCIENCE

In the 21st century, humanity has the hindsight to know that data science and its endeavours yield their own rewards. As a result, funding for the study of agricultural endeavours using data science is no longer too difficult to come by. The different areas under the umbrella that is agricultural data science have, in many ways, become fields unto themselves. Urban and precision farming, net-worked farms, and agricultural technological innovation all have their own sub-fields, but they all belong to the field of agriculture data science in one way or another. All of these different fields that make up modern agricultural data science have one attribute in common, however, and that one attribute is a focus on sustainability in whatever agricultural endeavour that is being undertaken. Indeed, modern data scientists and farmers would have a difficult time encouraging and practicing new processes that did not take sustainability into account. With the world’s population expected to keep growing, and the amount of food needed to feed the population expected to double by 2050 [13], this new focus on sustainability is an ever growing piece of the modern farming puzzle, whose importance and juxtaposition along side traditional and modern beliefs about agriculture can no longer afford to be overlooked. Data science factors into this tremendously as it enables farmers

and agriculturalists to analyze their data and figure out if their new approaches and techniques are actually working and worth continuing.

Focusing on sustainability and changing things around the farm are not the only new tricks that modern farmers and data scientists have up their sleeves. New techniques and advances in computing technology, both hardware and software, have helped pave the way towards making the analysis of farming data easier and more available than ever so that farmers and agriculturalists can make informed decisions about how they grow their crops and produce food. One example of a new technique that has been developed is based on an old approach: selective breeding. Selective and cross breeding different strains of plants is nothing new. However, thanks to modern technology and new techniques, farmers can work with data scientists to dig down to an even greater degree of analysis based on data that was not available in the past. One study's model, for example, analyzed how different nighttime temperatures and amounts of nitrogen fertilizer used impacted growth rates of rice. The study found that high night time temperatures "substantially reduces yields of cereal crops" [26]. Studies and experiments like this are specific examples of how agricultural data science is changing the ways farmers grow food. Because of studies like this, farmers are now breeding their rice strains selecting for traits that tolerate high nighttime temperatures. All of this data was obtained from sensors that were developed for the express purpose of gathering this kind of data. It is in this way that data science is encouraging agricultural advancement. But will improving the success rates of plants and a focus on sustainability be enough to help farmers provide food to so many people in the future? Time will tell, but data science is an incredibly useful tool when applied to producing food.

The focus on sustainability and hardier plants that produce more food, while an important part of the food puzzle, are not the only pieces that humanity needs to focus on in order to feed future populations. Architectural endeavours like vertical farms, designing and implementing a connected farms that take advantage of Internet of Things technology in farm equipment, using drone and wireless sensor monitoring systems technology, and using computing networks and sensors to figure out new information about insect and rodent infestation rates and crop losses are just a few examples of the kinds of outside the box thinking that is being done to improve agriculture. Through the use of technology and approaches developed from the use of data science practices, the agriculture sector as a whole continues to improve and is stepping up to meet the modern demands of an ever growing population.

### **3.1 Growing Urban Populations and the Greater Demand for Food**

More and more humans are beginning to live in centralized places like major urban cities while at the same time, people are leaving rural areas and communities in greater and greater numbers. This is having an impact on farming and agriculture in a number of ways. First and foremost, this is leading to a major reduction of the number of farmers and agriculture workers. According to the 2012 US Census of Agriculture data, less than 2% of the population of workers in the United States classify themselves as farmers - about

3.2 million people classify themselves as farmers, ranchers, or some other kind of agriculture related occupation [30]. This reduction in available farmers and agricultural workers means that the job of providing food to an ever growing population and society falls to fewer and fewer people. It also means that providing fresh fruits and vegetables to the growing and expanding urban populations is going to become more and more difficult due to the logistics of growing larger quantities of food, storing it, and then by getting it from point A to point B. One upside to this population shift is that the farms that still exist are getting much bigger to meet demands, and larger farms typically use more support tool and decision management data science tools. This means that more data than ever is being created, and more data is the cornerstone to using data science as a tool to improve agricultural production, efficiency, and sustainability.

[Figure 1 about here.]

[Figure 2 about here.]

Being able to produce food within urban areas is a dynamic approach to helping solve several problems. This process of growing food and other greenery in urban areas would not only help ease the demand for food from sources outside the urban areas, but it also helps to create less of an environmental impact while simultaneously promoting a sustainable model. By again causing even less strain on the farming resources outside of the urban sectors, this enables them to focus on a more manageable demand [21]. Growing what is needed locally within urban areas is a great way to help the environment, increase availability while also providing food and agricultural goods for a growing population, and encourage a sustainable agricultural model. One major benefit of farming in a close system like an urban farm is that since the system is closed, the urban farmers do not have to worry about insects, rodents, or other pest infestations that might destroy their crops. But more importantly, the nature of the closed system means they do not have to used insecticides, pesticides, or rodenticides. This means the urban crops are safer for consumption and that they leave a much smaller environmental foot print than their rural cousins [21]. This closed system method also allows the urban farmers to use the precise amount of resources that their data indicates they should be using to grow their plants. This increase in efficiency is not only an economic boost, but further improves this model of sustainability by making the environmental footprint of farming even smaller. Data science is also helping improve these endeavours in the same way it helped with farmers in the mid to late 20th century: by applying rigorously tested computing models to agricultural jobs, urban farmers are pushing the envelope in regards to how much food they can grow with limited spaces while saving on resources [1].

The United States and other developed nations are not the only places taking advantage of these advances in farming technology, newly developed processes for improving production, and taking advantage of agriculture data analysis. Developing nations are also benefiting from this new era where agriculture is meeting modern technology. As world populations grow and nations develop further, the demand for more and varied goods increases, including a demand for more varied foods. By taking advantage of data science practices, farmers and food providers in developing nations are

discovering new and innovative ways to meet this new demand placed on them for their goods. These new approaches are helping developed nations two fold: not only are they helping farmers to produce greater and more varied quantities of food, but they are also helping to limit the environmental footprint, created from farming, in places that are more sensitive to environmental change, or where environmental laws are not as heavily enforced [12]. The ability to meet food demands, curtail the effects of climate change on the surrounding environment, maximize the efficiency of resource use, and take advantage of advances in food storage and distribution are helping to transform developing nations in ways that all of their citizens can benefit from. Limiting and reducing the environmental footprint of farming and other agriculture processes is also extremely beneficial for developing countries and places that are facing more extreme, contemporary threats from climate change, as opposed to other nations whose economies and well being are less dependent upon their agriculture sector [12]. The direction that modern farmers in developing nations are taking, and their focus on sustainability, are only possible because data analysis tools have brought the world's agricultural expertise to this point. Keeping the data open and friendly allows for cross applicability of the data, which leads to more insights and discoveries, which in turn continues to benefit the farmers and agriculturalists even more.

### 3.2 Focusing On Sustainability

As mentioned previously, the growing focus on sustainability is helping to drive technological innovation and advancements and new techniques that produce more and better food while also limiting and reducing the environmental footprint required to do so. This is good news for farmers who are facing a shrinking population of agricultural workers in the face of growing demand for food in centralized urban areas. This growing demand has not gone unnoticed by the governments of the world, and many of them are actively taking steps by working with farmers and providing resources for research and development of farming practices that leave the ones farmers have been using in the dust. This focus on sustainability is not just for places like the United States who have the resources to explore new and dynamic avenues for agricultural experimentation. These new techniques and processes are also being quickly adopted by developing nations around the world in order to combat their countries' own hunger and resource problems. With the looming threat of climate change, whose impact is already beginning to make itself more and more apparent around the world, the demand and pursuit of data oriented precision agriculture is increasing at an exponential rate [29]. Since the well being of many developing nations is tied so closely with their agricultural production, they are the most susceptible to climate changes and the damage it can cause to food production [22]. Humanity is good at overcoming adversity, however, and data science is clearly helping to tackle the effects of climate change on agricultural endeavours and food production. The threat of climate change itself is driving entirely new agricultural fields whose sole focus is on sustainability.

There are many factors driving the technological innovation in data science focused agriculture. But of all of them, human caused climate change is perhaps one of the biggest factors driving the

changes and modern focus on sustainability, especially in developing nations. "Because most developing countries depend heavily on agriculture, the effects of global warming on productive croplands are likely to threaten both the welfare of the population and the economic development of the countries" [22]. Since developing nations are more sensitive to the effects of climate change because their economies and well being are often directly dependent on their agriculture sector, they are the ones who are benefiting the most from all the advancements in this field. These benefits are having a stabilizing effect in areas where these practices are being used, allowing for these places to develop further in areas that they ordinarily would not be able to focus on if they were still struggling to meet food requirements. By being able to focus on other parts of their society, these nations are able to further develop themselves and achieve greater and greater standards of living and freedoms for their citizens [20]. This is just one example of how agricultural data science has positive effects on society outside of the agricultural and environmental sectors. These effects, when used with noble intentions, are good for everyone.

### 3.3 Urban and Vertical Farming

As touched upon briefly earlier, when farmers begin producing some of the fruits and vegetables that people need inside of urban areas instead of on farmlands, this helps ease another one of the biggest problems farmers have encountered in the past, the problem of land availability. Again, as populations continue to grow, the stress they put on the demand for resources becomes more and more extreme. Land and resources becomes more and more scarce, not only because they are required for people to live on, but also because lots of other resources are required to handle large populations. Land resources for building roads for transportation, food and retail zones, resources such as land fills, waste removal, hazardous storage, water treatment, and power plants are just a few of the many other land resource demands that increase hand-in-hand as urban populations increase. One novel solution being used by many countries to tackle the problem of scarce geographic resources comes from thinking dynamically about the problem and realizing that, technically, farmland is not required in order to grow food and have a farm. Instead of expanding outward in order to grow more food, some modern farms are being rethought and built upwards or in repurposed, closed and controlled facilities in major urban areas. By utilizing or building multi-level structures laid out over a semi large area, farmers are able to grow different crops at different levels. These installations can be built in major urban areas, but any open urban space will do. This has the bonus side effect of reusing old buildings that might not have previously been in use anymore, which adds to and promotes a sustainable model. These structures allow for the same kind of closed system farming techniques that precision farming benefits from, while also allowing farmers to control everything that is done to their crops [11]. Having all the plants in urban areas also has the benefit of naturally cleaning the air. Plants use many of the gases released in vehicle emissions for their life functions. Taking these harmful gases out of the air is beneficial to humans, the plants, and the surrounding environment [11]. Building urban farms like this is a win for everyone involved, and as people and governments begin to take a more active role

in regards to improving and pursuing sustainable models of food production and environmental protection, urban farms are likely to gain in popularity and start popping up all around the world.

Another angle that can be taken in regards to vertical farming is the idea of growing plants on all available flat surfaces. Not only floors, but also ceilings and walls where available. One problem major urban areas can have is a lack of green spaces available. This takes away from the aesthetics of these urban locations, while also allowing for pollution to go to choke out major areas in cities. Growing plants on some walls and buildings around major cities will help reduce the impact of both of these problems on the people and their environment. The plants being around the city take care of the lack of green places on its own, transforming concrete jungles into lush, semi-green cities. Meanwhile, the plants themselves will help clean up pollutants in the air from human emissions and simultaneously reduce amounts of noise pollution in their immediate vicinity. These green walls can even be limited to urban agricultural buildings themselves and would still be effective and have a positive impact on their immediate environment [28] Again, data science makes all of this possible by allowing analysts and farmers to figure out the best ways to execute their agricultural endeavours, how to grow their plants, which and how much of their resources they need to use, and so forth. Technically, all of this could have and has been done in the past; it is not difficult to grow plants on the sides of buildings. But now urban populations are reaching heights that have never been seen, and the demand on environmental resources and human emissions are ever increasing. These simple approaches to the problems presented above are a means for cities to tackle a lot of the problems in city living, along with helping ease their dependency on farmlands for resources [28]. When urban farmers and city planners have access to data science and analysis tools that allow them to review and analyze information at a much deeper level, they are able to find new insights into the problems they are trying to solve. These new insights are driving urban farming to the level it needs to be at in order to meet the needs of an ever growing population and increased urban demand on resources.

### 3.4 Precision Farming: Networked Farms

The idea of a connected farm is paramount in moving beyond the historical approach to farming and agriculture. The cornerstone of understanding and finding better techniques to improve farming is data, and a networked, or smart farm, does just that. By using technology that networks the farmland, the farmer now has access to a decisions support network, which allows farmers the ability to keep track of all the happenings taking place on their property in ways they have not been able to in the past. This new attention to detail taking place allows farmers and agriculturalists to engage in a practice called precision agriculture. Precision agriculture “concentrates on providing the means for observing, assessing, and controlling agricultural practices” [17]. In essence, precision agriculture, or smart farming, focuses on sustainability and finding the sweet spot between resource use and crop growth and food production. Being able to hit that efficiency spot allows farmers to save on resource use while getting the most bang for their buck in regards to crops grown and sold per acre. Farmers are able to take advantage of all kinds of new data at their disposal. They

have access to modern technology, which allows them access to things like satellite telemetry data on not only the weather, but also insect populations and blooms, as well as a myriad of data on other farming techniques and practices that are still evolving to improve efficiency and production.

By taking advantage of new technology, as well as Internet of Things based technology, farmers and agriculturalists are able to tap into a source that humans have never been able to use in the past. Examples of modern agricultural technology include advances in wireless sensor technology that allow for the monitoring and changes in environmental conditions, resource use and precision agriculture, warehouse and storage management for storing crops and other perishables, technology that allows for large amounts of automation, and RFID technology that allows for tracking of the distribution of goods from farms [32]. All of these new technologies work in conjunction with one another to improve all aspects of the farm by making it possible to accurately monitor for and detect small problems that might arise and get them taken care of before they turn in to big problems. This proactive process of monitoring for problems fits with the growing importance of sustainability. Getting to problems and fixing them before they become larger issues can help the farm in countless ways. Practical ways in which data science technology can help improve farms are detecting insects, rodents, or other pests before they become an infestation, monitoring environmental conditions outside or in storage spaces in order to ensure that the crops they have stay fresh longer and do not become tainted in any way, determining the precise amount of resources required for individual plots of land or crops being grown so the farm’s resources are being used more efficiently, soil analysis to determine the best kinds of plant strains for their specific farmland, and tracking the distribution of their farm goods in order to more accurately distribute them to retailers. All of these agricultural techniques come together to form the larger picture that shows how much data science has really changed the agriculture sector. All of these technologies only came into existence in the last 25 years [32].

[Figure 3 about here.]

### 3.5 Modern Agricultural Data Science Beyond Food

Data science has the power to improve farming and agriculture in ways beyond just precision agriculture and growing food in ways and places that have never been done before. By taking a look at the entire picture, it is possible to shave even more off the proverbial top in terms of efficiency and improving sustainability in relation to farming and agriculture. Data science can improve the economic returns of local farmers while also helping to minimize the environmental footprint that is produced from the production and transportation of goods. Modern technology and machines enable for the harvesting, gathering, and preparation of food goods to be automated, faster, and much more efficient than if it was done by hand [32]. Modern technology also allows for the farmer and business owners to keep track of how much of their products are being sold and in which locations. This allows the retailer to order more precise amounts to fit their needs while also informing the farmer which crops are best to grow, when to grow them, and what

quantities to shoot for. GPS and other transportation technology can be used in the transportation process to make sure that the drivers have the most direct and efficient routes possible while delivering their goods, and advances in communications technology make it easy for orders to be changed or updated at the last minute [29]. The goal of all of this is to produce less overall waste and put less stress on the environment. Data science helps mitigate problems that would produce more waste and add stress to the environment, so in this way, it is one of the most important tools humanity has in solving these problems and continually improving the agriculture sector to meet the demands of bigger populations.

### 3.6 Setbacks and Steps Towards the Future

The ability to analyze agriculture data with modern technology is leading to many unexpected discoveries. With sustainability and combating climate change being two of the most important driving factors in innovation, scientists and farmers are finding are using data science models to find dynamic solutions to problems, both old and new. At this point one of the biggest problems holding back agricultural data science, despite the fact that more and more data is being generated everyday, is the distinct lack of data. There are many challenges that must be overcome as we move towards the future of agriculture, but one of the greatest obstacles “is to obtain reliable data on farm management decision making, both for current conditions and under scenarios of changed bio-physical and socio-economic conditions” [6]. In other words, it is not a question of having reliable data, as much as it is a question of having reliable data that pertains to scenarios and circumstances that are difficult to reproduce, that have not happened yet, but are speculated to happen as the climate change. Despite this, the modern data that has been collected is obviously still being put to good use and helping to solve major problems that humanity knows will be immense obstacles in the future.

One example of a modern solution to an old problem is fighting emissions that pollute the earth. Although it is often thought that vehicle and airplane emissions cause the most air pollution, but of all of humankind’s endeavours, it is factory farming that is having the biggest impact on our environment and exacerbating the effects of climate change [7]. Agricultural data science is being used to help combat the effects of factory farming in a number of unorthodox ways. One recent example of agricultural data science making a breakthrough in this area came when scientists discovered that feeding cows, who are by far the biggest producers of methane and other remissions, ground up bits of seaweed with their regular feed will radically reduce the amount of methane they produce while having no negative affect on the animals [19]. The wireless monitoring sensors and models used to ascertain these findings were only available because they were created from investments in the pursuit of agricultural data science practices. As findings like this become more common and see widespread adoption, the environmental footprint of factoring farming will begin to decline. This will be incredibly useful for developed nations whose factoring farming emissions levels are continuing to rise [7].

[Figure 4 about here.]

## 4 THE FUTURE OF AGRICULTURAL DATA SCIENCE

The future of agricultural data science is concerning itself with not only continuing to solve and improve the same old problems, but also exploring entirely new, out of this world concepts in regards to farming and growing foods. In the past, agricultural data science was focused on gathering data and growing the field. Modern agricultural data scientists are taking on problems like climate change, staggering populations and their demand for food, and finding new ways to improve sustainable agricultural models. So, then, it seems that the future of agricultural data science is beginning to come into focus. Although new problems and fields are bound to arise, the focus of the future of agricultural data science seems to be on automation and enhanced sustainability through disturbing the environment as little as possible [18]. Since machines are able to perform tasks more efficiently than humans can, reaching a point of agricultural automation is one of the potential goals of sustainable models. A mostly automated farm is much more efficient than one that relies on the human element to perform jobs and work. That is not to say that there will never be a human element involved, but the future farmers may have more in common with data scientists and programmers than they do with their modern and historical counterparts who worked in fields most of the day [23]. As technology improves in ways we cannot imagine right now, the possibilities of how data science will influence agriculture in the future are great.

The future may seem like science fiction in many ways, but modern technology and agricultural procedures may have seemed like science fiction if they were explained to a farmer from the 1950s. That does not mean that we are not able to see where a lot of different areas or advancing, to theorize technology and procedures that farmers and agriculturalists are not able to take advantage of now because they are too expensive. But as technology becomes cheaper and new methods and data science models are built, the future quickly becomes the present as humans catch up with their imaginations. Advances in automation, bloodless food production, extra-solar farming, and eventually terraforming, when realized, have the potential to transform our society in astonishing ways, possibly even leading to a post scarcity society where everyone’s basic needs are met. Where agriculture itself enabled humankind to stop focusing on strictly survival and evolve into societies, so too might automated agriculture and food production allow for humanity to achieve a new level of societal evolution [10].

The coming changes in agricultural data science are not simply limited to technological or physical. Indeed, even now as humanity’s understanding of its impact on the climate and surrounding environment is coming to be better understood, world governments are beginning to realize the importance of sustainability and limiting the environmental footprint that is created from food production. World governments taking an active interest is having a positive effect on the research and development being done in the fields related to agricultural data science [6]. This new emphasis is changing the way many politicians think about agriculture and making them eager to use political leverage to enact changes in government which put guidelines into place and make resources available

that enable agricultural researches to analyze their data and make informed decisions when attempting new agricultural endeavours.

#### 4.1 The Future and Automation

Much as sustainability and environmental impact are big tent poles driving innovation in agricultural data science and technology, so too will they continue to be in the future. However, a third piece of the puzzle is being thought of as more data is generated and analyzed, and that piece is automation. Automation is when technology and machines work to perform the jobs and tasks that humans would ordinarily do. Automation makes work far more efficient because, typically, a machine requires less resources to perform work than a human does. The resources a machine requires come largely in the form of costs upfront, but they quickly pay for themselves [18]. Automation allows for producers to produce their product around the clock, so too would the crops on an automated farm receive constant care and attention that a human would be unable to provide. This fine attention to detail would increase the amount of food produced and improve resource use efficiency for growing food.

This concept is not referring to an enormous, single intelligent farm that knows all. Instead, an automated smart farm would actually be a collection of many smaller, automated systems that all work together to ensure the success of the farm and its food production requirements. Not all farms would need to be completely autonomous in order to benefit from this technology. For example, some farms are already taking advantage of available technology to automate simple tasks and jobs that humans used to perform, such as automated irrigation systems [14] and tools that automatically extract key data bits from current crop conditions and execute automated commands to tend the farm, based on a set of predefined variables [9]. Nonetheless, as more and more automated systems are made available and become less expensive, more farms will be adopting them, leading to an even greater level of automation and requiring even fewer workers with diverse skill sets than ever before.

#### 4.2 Meat Grown in Labs

Climate change is one of the biggest threats spurring research and interest in efforts to improve sustainable agriculture practices. As mentioned earlier, the one human activity that by far has the greatest impact on the environment is factory farming, which is a process of raising cattle and other livestock in controlled conditions [7]. Similar in approach to precision agriculture, factory farming in developed nations is having a substantial impact on the environment from all the emissions the animals produce, as well as the resources that are required to operate factory farming facilities. This poses a problem for future generations since it is an unsustainable model. One solution being aided by data science that is currently too expensive is the possibility of growing meat in a laboratory setting without requiring the growing and slaughtering of actual animals. Basically, this process is “a novel idea of producing meat without involving animals with the help of tissue engineering techniques” [2]. At present levels of technology, this process is possible but far too expensive to be a practical solution to producing food on a large scale. However, data science models are helping to drive

down costs by allowing scientists to analyze their data and find more efficient ways to accomplish their goals. Once it becomes less expensive to grow meat in a lab that is indistinguishable from traditional meat, protein farms will likely start popping up all over the world [2].

There are other benefits to think about when considering the impact of switching from traditional meat production to lab grown. The biggest, as mentioned slightly above, is that it will help to dramatically reduce the environmental footprint being created by factory farming practices. Growing meat in labs, once costs and techniques are worked out, has the potential to be radically more sustainable model than humankind’s current practices [15]. Once it becomes possible and feasible to be able to grow all the healthy meat needed to satisfy the demands of the growing population in a lab, this model will begin to be adopted because of the economic and environmental advantages to its use that come with following a model focused on sustainability. Data science is helping to make this more of a reality by producing more advanced models that help scientists and engineers get their jobs done and find newer, cheaper ways to produce this lab grown meat [15].

[Figure 5 about here.]

#### 4.3 Farming in Space

One undeniable truth about humanity is that humans expanding and exploration seem to be hardwired into our genetic makeup. Since the beginning of humankind’s history, exploration and settlement have been a big part of what drives human innovation. Necessity is the mother of invention, as the saying goes. It is in the spirit of that saying that future agricultural endeavours are being theorized and planned for today. One big possibility on the horizon is the necessity to grow food and farm off-world because of lack of space, resources, or environmental factors making the available farmland incapable of meeting the demands of future populations [23]. Once agricultural development and food production reach a tipping point in regards to demands and human population, there will be no choice but to start farming in space. Enormous and fantastical space stations could be constructed where food could be grown in closed systems. This endeavour, though expensive, would eventually pay for itself and allow for a level of control, efficiency, and automation not available on Earth [23]. By designing and constructing a station like this from the ground up, with things like sustainability and efficiency in mind, food will be able to be produced in a way never before practiced by humanity. This has the added benefit of having absolutely no impact on the environment, since it is not even being done on Earth. Right now, space travel and getting super structures like this built are prohibitively expensive. However, the costs of such things are expected to go down as technology advances and methods of space travel become more available [31].

#### 4.4 Terraforming Worlds: the Height of Agriculture

At the most extreme end of outside the box thinking comes the most fantastic sounding concept yet: terraforming. Terraforming is the process of making another planet or heavenly body habitable for humans to live and thrive on. In the distant future there may come

a time when humanity needs to take steps to become a multi-planet species. Data science will be invaluable when achieving this level of agricultural endeavour as the amount of data to be processed and understood will require data analysis models and techniques that have not even been invented yet [8]. The ability to adapt a planet to human life would require a complete mastery of agriculture which could only be obtained through refined understanding of unimaginably large amounts of data created from attempting such a task. Right now, scientists can only re-create extraterrestrial soil in a labs and perform experiments to grow food there, so any work done in this field is mostly hypothetical, but not outside the realm of possibility. If humans continue to advance at the exponential pace at which they are, one thing begins to become undeniable clear. Given a long enough timeline, environmental, societal, or geographical conditions will come about that will one day make it a necessity for humans to start living on multiple planets. Although this is science fiction now, data science and the insights its use grants us are making the impossible possible everyday.

One applicable experiment that was performed to test the validity of adapting the soil on Mars to growing terrestrial flora was when scientists looked at the recent volcanic iron deposits in Santorini, Greece. The bizarre lifeforms that were found living in this environment “provides a potentially useful ecosystem for Mars terraforming experiments” [24]. Using data science to gather and make sense of the data generated on terrestrial locations that are similar to extraterrestrial ones brings humanity one step closer to terraforming, even if it is a baby step. As fantastical a concept as bending another planet to humanity’s will is, if you stop to think about it, this is nothing new. Humans have been terraforming the Earth for thousands of years already, just not in the ways we would prefer. Terraforming on a large scale is theoretically possible, but it will never be possible without the data science tools and techniques to analyze the mountains of data that would be need to be analyzed to achieve such a advantageous goal that may one day be a necessity.

#### 4.5 Restructuring Society

The possibility of achieving a post-scarcity society, while seemingly outlandish considering humanity’s current problems, could become a reality in the future. Having all of humanity’s food needs met automatically through space aged inventions like massive orbiting space farms and home or lab grown meat would lead to a restructuring of society humans have not seen since we first started farming twelve thousand years ago [10]. Should humanity ever achieve this level of societal progress, data science methods and models will be largely to thank for allowing humans to understand and improve their work by analyzing the data it produces. A largely automated society would produce an enormous amount of data to be analyzed, which as previously discussed, has the benefit of becoming even more sophisticated as more data is generated to learn from [3]. This self reinforcing system of generating data, improving from it, then generating more is showing no signs of slowing down as humanity is only just now beginning to see the benefits of complex automation. Self driving vehicles and automated farms are on the horizon, as well as a myriad of other technological innovations, and data

science and its versatile applications are one of the biggest reasons that humanity has to thank for these technological possibilities.

### 5 CONCLUSION

Humanity’s modest roots as simple nomads who discovered that they could grow their own food and use agriculture as a tool to build civilizations lasted for thousands and thousands of years. There were advancements, sure, but they were slow in coming and lacking in sophistication. However, in the last 75 years or so, humanity has seen the agriculture sector explode with new developments in techniques, technologies, and practices that have increased food production and allowed the agricultural sector to keep pace with growing populations that have a greater demand for food and agricultural resources. How was the agricultural sector able to revolutionize itself when in the past, changes came about much more slowly and incrementally? The answer, of course, is that the field of data science has been one of the chief tools used to solve agricultural problems and find solutions that meet the demands of today.

Technological advancements in hardware and modeling systems are ushering in a whole new era of human agriculture that focuses on sustainability. Producing as little waste as possible, while also impacting the environment in ways that contribute to climate change in as few of ways as they can are among the most important goals of modern day agriculture and food production, and data science is helping farmers and agriculturalists achieve these beefy goals. By creating vast networked farms, modern day farmers and agriculturalists have access to a level of data analysis that has not been available in the past, enabling them to make informed decisions and change the way they do things in order to improve the efficiency and output of their farms. This analysis is also leading to improved sustainability practices on farms by allowing the farmers to understand statistics about resource use and relation to crop growth like they could not in the past. Urban farming is also proving to be a modern solution to food distribution problems in highly populated areas, while also having the beneficial side effect of the plants helping to clean and detoxify the potentially harmful human made emissions that are found in major cities around the world.

The field of data science, from its origin to its current state as one of the premiere methods of human problem solving, is only going to continue to become more and more sophisticated as time goes on. With computer technology continuing to become smaller, cheaper, and more powerful, as well as data analysis models increasing in their reach and sophistication, the future of decision management tools and informed decision making that is going to be available to farmers and agricultural workers will be staggering. The future that agricultural data science is enabling seems like science fiction, but it is rapidly becoming a reality. Major projects like orbiting farms that meet the demands of the Earth’s people to terraforming entire planets are going to require advanced tools and models that only sophisticated data science techniques and models will be able to provide in the future. Although humanity is still going through the growing pains of becoming a global, connected community, the future looks bright, statistically speaking. World hunger is still a problem humans are trying to solve, but data science has empowered us to fight it on a level battlefield where. Assuming

human innovation continues at the exponential rate it has set for itself, the struggle to feed every human is a battle data science will help us to win.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Miao Jiang, and Juliette Zerick for assistance with my assignments and using github.

## REFERENCES

- [1] Hermann Auernhammer. 2001. Precision farming – the environmental challenge. *Computers and Electronics in Agriculture* 30, 1-3 (feb 2001), 31–43. [https://doi.org/10.1016/s0168-1699\(00\)00153-8](https://doi.org/10.1016/s0168-1699(00)00153-8)
- [2] Zuhail Fayaz Bhat, Sunil Kumar, and Hina Fayaz Bhat. 2015. In vitro meat: A future animal-free harvest. *Critical Reviews in Food Science and Nutrition* 57, 4 (may 2015), 782–789. <https://doi.org/10.1080/10408398.2014.924899>
- [3] Alain Biem, Maria A. Butrico, Mark D. Feblowitz, Tim Klinger, Yuri Malitsky, Kenney Ng, Adam Perer, Chandra Reddy, Anton V. Riabov, Horst Samulowitz, Daby Sow, Gerald Tesauro, and Deepak Turaga. 2015. Towards Cognitive Automation of Data Science. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, Austin, Texas, Article AAAI'15, 2 pages. <http://dl.acm.org/citation.cfm?id=2888116.2888360>
- [4] Norman E. Borlaug. 2002. Feeding a world of 10 billion people: The miracle ahead. *In Vitro Cellular & Developmental Biology - Plant* 38, 2 (mar 2002), 221–228. <https://doi.org/10.1079/ivp2001279>
- [5] Norman E. Borlaug. 2007. Sixty-two years of fighting hunger: personal recollections. *Euphytica* 157, 3 (jun 2007), 287–297. <https://doi.org/10.1007/s10681-007-9480-9>
- [6] Susan M. Capalbo, John M. Antle, and Clark Seavert. 2017. Next generation data systems and knowledge products to support agricultural producers and science-based policy decision making. *Agricultural Systems* 155 (jul 2017), 191–199. <https://doi.org/10.1016/j.agsy.2016.10.009>
- [7] Dario Caro, Steven J. Davis, Simone Bastianoni, and Ken Caldeira. 2014. Global and regional trends in greenhouse gas emissions from livestock. *Climatic Change* 126, 1-2 (jul 2014), 203–216. <https://doi.org/10.1007/s10584-014-1197-x>
- [8] Amber Dance. 2016. Science and Culture: Terraforming a volcano, artfully. *Proceedings of the National Academy of Sciences* 113, 16 (apr 2016), 4234–4235. <https://doi.org/10.1073/pnas.1603563113>
- [9] Jnaneswar Das, Gareth Cross, Chao Qu, Anurag Makineni, Pratap Tokekari, Yash Mulgaonkar, and Vijay Kumar. 2015. Devices, systems, and methods for automated monitoring enabling precision agriculture. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, Austin, TX, Article 10.1109/coasee.2015.7294123, 12 pages. <https://doi.org/10.1109/coasee.2015.7294123>
- [10] Matthew David. 2017. Sharing: post-scarcity beyond capitalism? *Cambridge Journal of Regions, Economy and Society* 10, 2 (feb 2017), 311–325. <https://doi.org/10.1093/cjres/rxs003>
- [11] Dickson Despommier. 2013. Farming up the city: the rise of urban vertical farms. *Trends in Biotechnology* 31, 7 (jul 2013), 388–389. <https://doi.org/10.1016/j.tibtech.2013.03.008>
- [12] T. Garnett, M. C. Appleby, A. Balmford, I. J. Bateman, T. G. Benton, P. Bloomer, B. Burlingame, M. Dawkins, L. Dolan, D. Fraser, M. Herrero, I. Hoffmann, P. Smith, P. K. Thornton, C. Toumlin, S. J. Vermeulen, and H. C. J. Godfray. 2013. Sustainable Intensification in Agriculture: Premises and Policies. *Science* 341, 6141 (jul 2013), 33–34. <https://doi.org/10.1126/science.1234485>
- [13] Rhys E. Green, Stephen J. Cornell, Jörn P. W. Scharlemann, and Andrew Balmford. 2005. Farming and the Fate of Wild Nature. *Science* 307, 5709 (2005), 550–555. <https://doi.org/10.1126/science.1106049>
- [14] Joaquin Gutierrez, Juan Francisco Villa-Medina, Alejandra Nieto-Garibay, and Miguel Angel Porta-Gandara. 2014. Automated Irrigation System Using a Wireless Sensor Network and GPRS Module. *IEEE Transactions on Instrumentation and Measurement* 63, 1 (jan 2014), 166–176. <https://doi.org/10.1109/tim.2013.2276487>
- [15] Olive Heffernan. 2017. Sustainability: A meaty issue. *Nature* 544, 7651 (apr 2017), S18–S20. <https://doi.org/10.1038/544s18a>
- [16] James W. Jones, John M. Antle, Bruno Basso, Kenneth J. Boote, Richard T. Conant, Ian Foster, H. Charles J. Godfray, Mario Herrero, Richard E. Howitt, Sander Janssen, Brian A. Keating, Rafael Munoz-Carpenna, Cheryl H. Porter, Cynthia Rosenzweig, and Tim R. Wheeler. 2017. Brief history of agricultural systems modeling. *Agricultural Systems* 155 (jul 2017), 240–254. <https://doi.org/10.1016/j.agsy.2016.05.014>
- [17] Mohamed Rawidean Mohd Kassim, Ibrahim Mat, and Ahmad Nizar Harun. 2014. Wireless Sensor Network in precision agriculture application. In *2014 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, Austin, TX, Article 10.1109/cits.2014.6878963, 9 pages. <https://doi.org/10.1109/cits.2014.6878963>
- [18] Michael Kassler. 2001. Agricultural Automation in the new Millennium. *Computers and Electronics in Agriculture* 30, 1-3 (feb 2001), 237–240. [https://doi.org/10.1016/s0168-1699\(00\)00167-8](https://doi.org/10.1016/s0168-1699(00)00167-8)
- [19] Robert D. Kinley, Rocky de Nys, Matthew J. Vucko, Loreenna Machado, and Nigel W. Tomkins. 2016. The red macroalgae Asparagopsis taxiformis is a potent natural antimethanogenic that reduces methane production during in vitro fermentation with rumen fluid. *Animal Production Science* 56, 3 (2016), 282. <https://doi.org/10.1071/an15576>
- [20] David R. Lee. 2005. Agricultural Sustainability and Technology Adoption: Issues and Policies for Developing Countries. *American Journal of Agricultural Economics* 87, 5 (nov 2005), 1325–1334. <https://doi.org/10.1111/j.1467-8276.2005.00826.x>
- [21] F Martellozzo, J-S Landry, D Plouffe, V Seufert, P Rowhani, and N Ramakutty. 2014. Urban agriculture: a global analysis of the space constraint to meet urban vegetable demand. *Environmental Research Letters* 9, 6 (2014), 064025. <http://stacks.iop.org/1748-9326/9/i=6/a=064025>
- [22] R. Mendelsohn and A. Dinar. 1999. Climate Change, Agriculture, and Developing Countries: Does Adaptation Matter? *The World Bank Research Observer* 14, 2 (aug 1999), 277–293. <https://doi.org/10.1093/wbro/14.2.277>
- [23] L. Purdy. 2016. Farming from space. *Engineering & Technology* 11, 2 (mar 2016), 40–44. <https://doi.org/10.1049/et.2016.0203>
- [24] Eleanor Iberall Robbins, Chrysoula Kourtidou-Papadeli, Arthur S. Iberall, Gordon L. Nord, and Motoaki Sato. 2015. From Precambrian Iron-Formation to Terraforming Mars: The JIMES Expedition to Santorini. *Geomicrobiology Journal* 33, 7 (sep 2015), 1–16. <https://doi.org/10.1080/01490451.2015.1074322>
- [25] David C. Rose, William J. Sutherland, Caroline Parker, Matt Lobley, Michael Winter, Carol Morris, Susan Twining, Charles Ffoulkes, Tatsuya Amano, and Lynn V. Dicks. 2016. Decision support tools for agriculture: Towards effective design and delivery. *Agricultural Systems* 149 (nov 2016), 165–174. <https://doi.org/10.1016/j.agsy.2016.09.009>
- [26] Wanju Shi, Gui Xiao, Paul C. Struiik, Krishna S.V. Jagadish, and Xinyou Yin. 2017. Quantifying source-sink relationships of rice under high night-time temperature combined with two nitrogen levels. *Field Crops Research* 202 (feb 2017), 36–46. <https://doi.org/10.1016/j.fcr.2016.05.013>
- [27] Neil Stephens and Martin Ruijkamp. 2016. Promise and Ontological Ambiguity in the in vitro Meat Imagescape: From Laboratory Myotubes to the Cultured Burger. *Science as Culture* 25, 3 (jul 2016), 327–355. <https://doi.org/10.1080/09505431.2016.1171836>
- [28] Suparwoko and Betri Taufani. 2017. Urban Farming Construction Model on the Vertical Building Envelope to Support the Green Buildings Development in Sleman, Indonesia. *Procedia Engineering* 171 (2017), 258–264. <https://doi.org/10.1016/j.proeng.2017.01.333>
- [29] A Trauger. 2009. Social agency and networked spatial relations in sustainable agriculture. *Area* 41, 2 (jun 2009), 117–128. <https://doi.org/10.1111/j.1475-4762.2008.00866.x>
- [30] "U.S. Census Bureau". 2014. 2012 Census. U.S. Department of Agriculture. (May 2014).
- [31] Maria Antonietta Viscio, Eugenio Gargioli, Jeffrey A. Hoffman, Paolo Maggiore, Andrea Messidor, and Nicole Viola. 2014. A methodology for innovative technologies roadmaps assessment to support strategic decisions for future space exploration. *Acta Astronautica* 94, 2 (feb 2014), 813–833. <https://doi.org/10.1016/j.actaastro.2013.10.004>
- [32] Ning Wang, Naiqian Zhang, and Maohua Wang. 2006. Wireless sensors in agriculture and food industry—Recent development and future perspective. *Computers and Electronics in Agriculture* 50, 1 (jan 2006), 1–14. <https://doi.org/10.1016/j.compag.2005.09.003>

#### LIST OF FIGURES

1	The decline of US farmers [30].	12
2	Farm occupation statistics [30].	12
3	Projections show how quickly the technology was growing at the start of the 21st century [32].	13
4	Cattle emissions rate trends over the years [7].	13
5	Simplified process by which stem cells grow edible meat [27].	14

## Number of U.S. Farmers, 2007 and 2012

<b>Operators</b>	<b>2007</b>	<b>2012</b>	<b>% change</b>
Principal	2,204,792	2,109,303	-4.3*
Second	931,670	928,151	-0.4
Third	145,072	142,620	-1.7
All	3,281,534	3,180,074	-3.1

\*Statistically significant change.

Source: USDA NASS, 2012 Census of Agriculture.

Figure 1: The decline of US farmers [30].

### Gender, Primary Occupation, and Years on Farm, 2012 (percent)

Farm Operators	Gender		Primary Occupation		Years on Farm	
	Male	Female	Farm	Other	<10	10+
Principal	86	14	48	52	22	78
Second	33	67	37	63	31	69
Third	61	39	43	57	45	55
All	70	30	44	56	26	74

Source: USDA NASS, 2012 Census of Agriculture.

Figure 2: Farm occupation statistics [30].

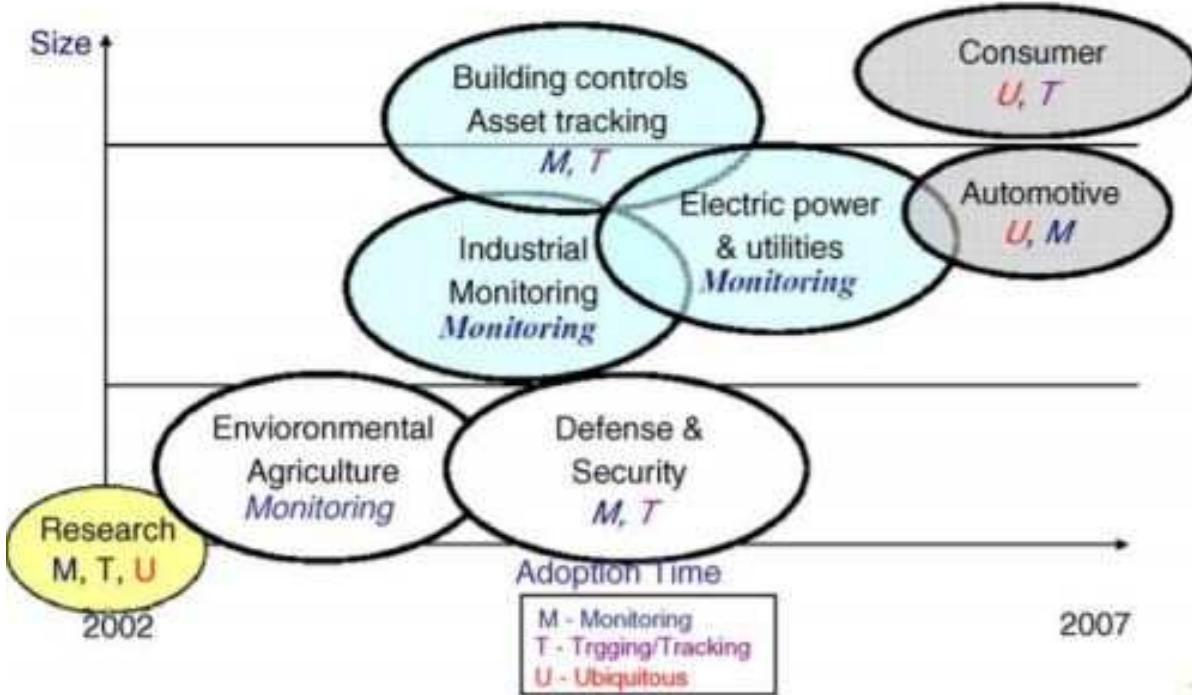


Figure 3: Projections show how quickly the technology was growing at the start of the 21st century [32].

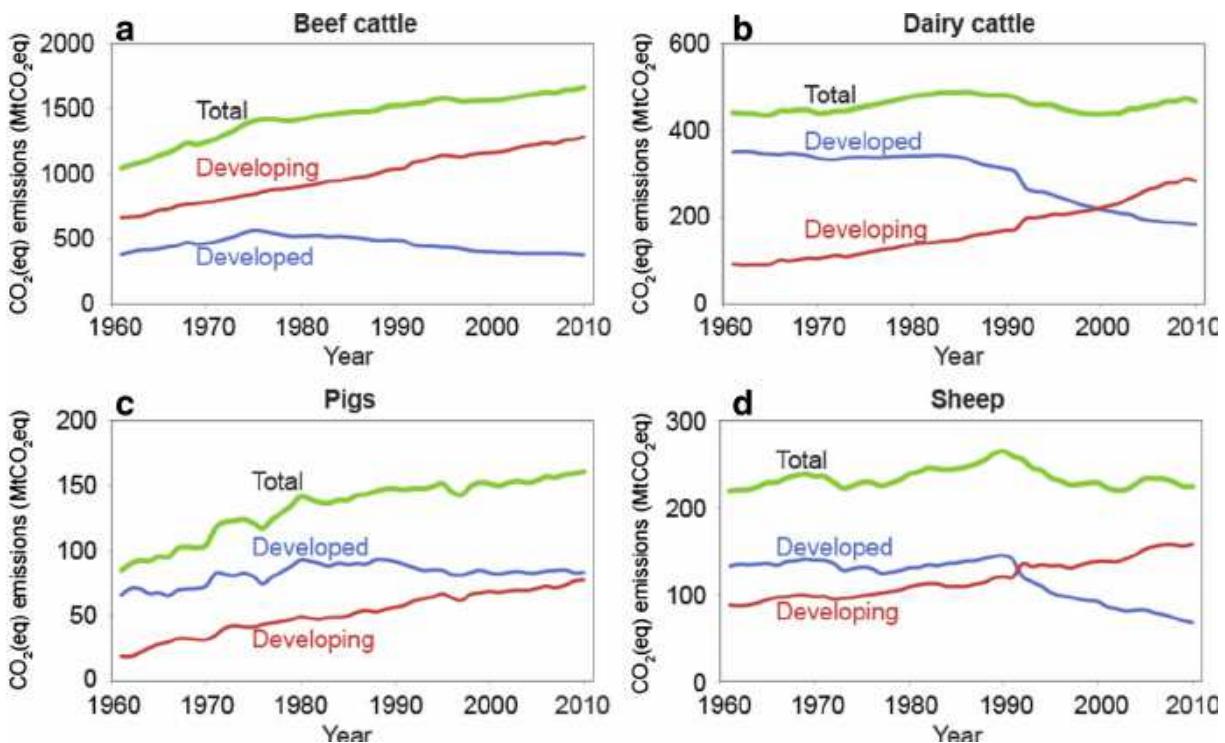


Figure 4: Cattle emissions rate trends over the years [7].

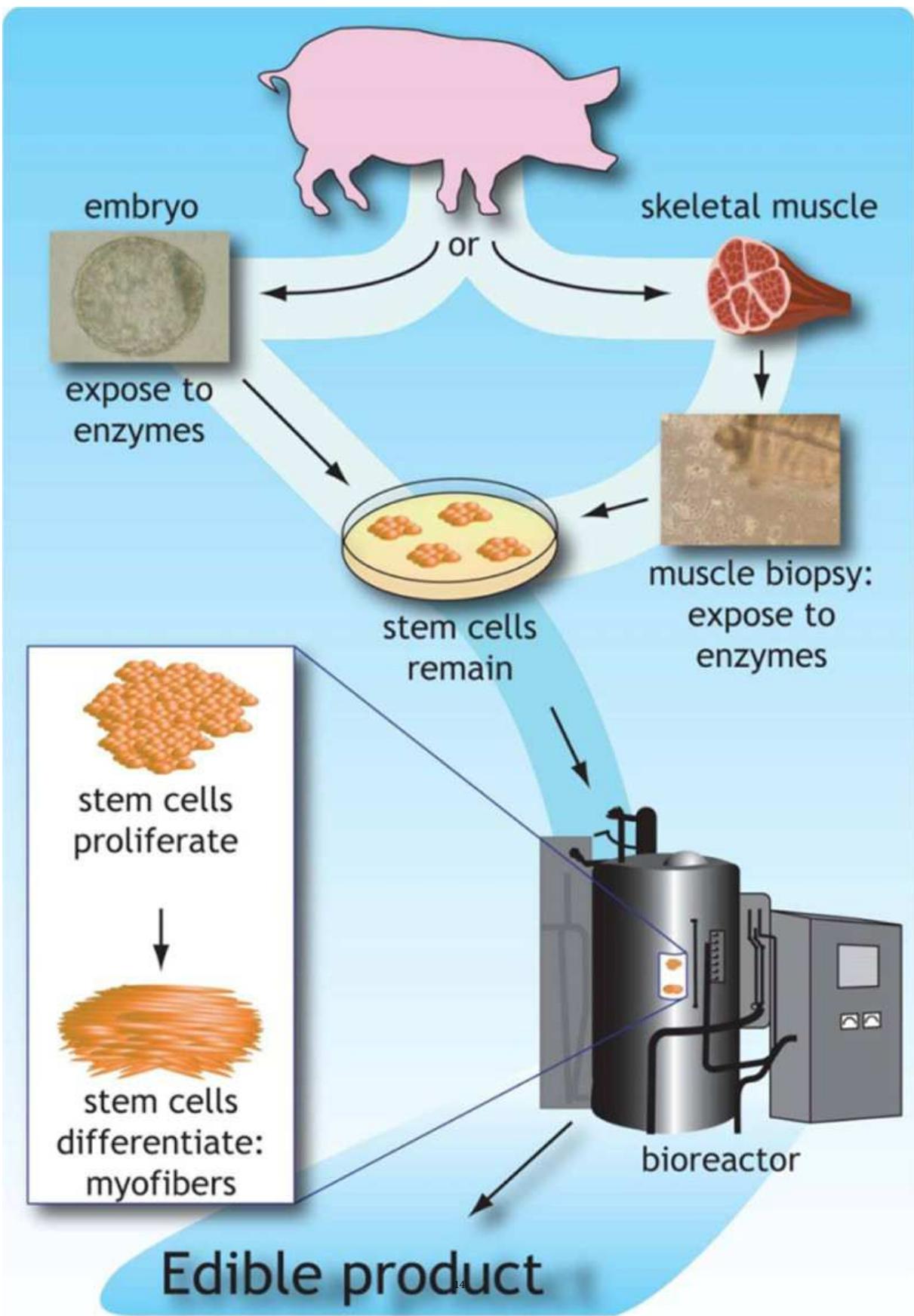


Figure 5: Simplified process by which stem cells grow edible meat [27].

```
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 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Ibera
while executing---line 3085 of file ACM-Reference-Format.bst
Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Ibera
while executing---line 3085 of file ACM-Reference-Format.bst
Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Ibera
while executing---line 3229 of file ACM-Reference-Format.bst
Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Ibera
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 4 error messages)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-12-10 13.53.10] 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 49.44307pt.
Typesetting of "report.tex" completed in 1.2s.
```

---

```
Compliance Report
```

---

```
name: Ross Wood
hid: 345
paper1: 100% re-review on 10/25/17
paper2: 100%
project: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
14
wc 345 project 14 9943 report.tex
wc 345 project 14 10593 report.pdf
wc 345 project 14 2295 report.bib

find "
-----
passed: True

find footnote
-----
passed: True

find input{format/i523}
-----
4: \input{format/i523}

passed: True

find input{format/final}
-----
passed: False

floats
-----
71: \begin{figure}
72: \includegraphics[width=.9\columnwidth]{images/fig1.png}
76: \begin{figure}
77: \includegraphics[width=.9\columnwidth]{images/fig2.png}
103: \begin{figure}
104: \includegraphics[width=.9\columnwidth]{images/fig3.png}
118: \begin{figure}
119: \includegraphics[width=.9\columnwidth]{images/fig4.png}
143: \begin{figure}
144: \includegraphics[width=.9\columnwidth]{images/fig5.png}

figures 5
tables 0
includegraphics 5
labels 0
```

```
refs 0
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
```

---

```
WARNING: figure and above may be used improperly
```

95: Another angle that can be taken in regards to vertical farming is the idea of growing plants on all available flat surfaces. Not only floors, but also ceilings and walls where available. One problem major urban areas can have is a lack of green spaces available. This takes away from the aesthetics of these urban locations, while also allowing for pollution to go to choke out major areas in cities. Growing plants on some walls and buildings around major cities will help reduce the impact of both of these problems on the people and their environment. The plants being around the city take care of the lack of green places on its own, transforming concrete jungles into lush, semi-green cities. Meanwhile, the plants themselves will help clean up pollutants in the air from human emissions and simultaneously reduce amounts of noise pollution in their immediate vicinity. These green walls can even be limited to urban agricultural buildings themselves and would still be effective and have a positive impact on their immediate environment \cite{suparwoko2017} Again, data science makes all of this possible by allowing analysts and farmers to figure out the best ways to execute their agricultural endeavours,

how to grow their plants, which and how much of their resources they need to use, and so forth. Technically, all of this could have and has been done in the past; it is not difficult to grow plants on the sides of buildings. But now urban populations are reaching heights that have never been seen, and the demand on environmental resources and human emissions are ever increasing. These simple approaches to the problems presented above are a means for cities to tackle a lot of the problems in city living, along with helping ease their dependency on farmlands for resources \cite{suparwoko2017}. When urban farmers and city planners have access to data science and analysis tools that allow them to review and analyze information at a much deeper level, they are able to find new insights into the problems they are trying to solve. These new insights are driving urban farming to the level it needs to be at in order to meet the needs of an ever growing population and increased urban demand on resources.

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 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Iberall" while executing---line 3085 of file ACM-Reference-Format.bst

Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Iberall" while executing---line 3085 of file ACM-Reference-Format.bst

Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Iberall" while executing---line 3229 of file ACM-Reference-Format.bst

Name 4 in "Eleanora Iberall Robbins and Chrysoula Kourtidou-Papadeli and Arthur S. Iberall" while executing---line 3229 of file ACM-Reference-Format.bst

(There were 4 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
```

# Big Data Analysis for Wild File Prevention, and Tracking

Zachary Meier  
Indiana University  
Bloomington, Indiana 47408  
zrmeier@indiana.edu

## ABSTRACT

In Southern California, wildfires never seem too out of place and seem to be a commonly accepted as a norm for the most part. Over the years this problem has only continued to get worse. These fires started to cause massive amounts of property damage, loss of life, and hazardous conditions for everyone. When wildfires break out they soon become unpredictable due to wind conditions and vegetation. To amplify problems further, most places where burns take place are hard to reach by the emergency crews fighting them. The combination of these factors results in scattered resources, playing catch up rather than preventing the spread and evacuating large areas. Using Predictive analytics to not only know where a fire is going but, possibly preventing it in the first place may help save lives and billions of dollars.

## KEYWORDS

hid346, i523

## 1 INTRODUCTION

Wildfires are a large part of life in Southern California, and just like any other type of event has a large effect on the population, we as humans do our best to prepare for it. Originally, it was building houses with stucco and tile roofs so they were less likely to catch fire. Now with the advancement of big data, there is a new way to look at and defend against wildfires. These new technology and methods will allow first responders to be ahead of the fire. Allowing for them to set backfires, and protect properties more efficiently, as well as help prevent new fires popping up from Santa Ana winds which help carry fire embers to new locations. There are a couple new technologies that help deal with this exact type of problem.

## 2 THE DATA

Data, the gold of the information era we currently live in. However, in some ways it is a much more complex item to mine for. It is ubiquitous, which end up adding more layers to the complexity of squiring the gold we are after. To fine the information we are after we need to understand what we are looking for. Most times it is something already in place, and needs no new real creation to squire it. Only by manipulating this data, can we get the answers we need. Below will give a brief insight into the data and framework usage. Then we can start to look at how programs are being built to harness this data in new, creative, and useful ways.

### 2.1 Capturing Fire

For years fire has always been seen as something natural and unpredictable. The same we would view a wild animal. There is a certain mystery to it yet it is a force that demands respect. Though the respect of this powerful force of nature may never dwindle.

The mysteries of how it moves, burns and it's overall behavior are becoming less so. To know where fire is going to go, and to simulate it correctly you need two types of data. There is a need for static data, and a need for dynamic data. These two types are dependent. If static data were to be used only, we would know what a fire would like to burn most likely, but that is about it. However if you layer dynamic data into that situation, you create many possibilities that mirror real life conditions, and allow for fire fighters to, know where the fire is going to spread.

### 2.2 Static Data

The static data that will most likely be seen in this instance are maps. The understanding of vegetation in the geographic area allows of an understanding of what is most likely to fuel the fires spread. In addition there are topographic maps which show elevation of the land, which can also help understand how a fire might move through different terrains.

### 2.3 Dynamic Data

This is the information captured that can help understand what affects fire behavior and spread in a certain amount of time. This data for wild fires can be broken up in to two different categories, Graphical User Interface (GIS) fire perimeters, and meteorological variables.

*2.3.1 User Interface Fire Perimeters.* In order to predict the fire and its behavior there is a need of accurately identifying where it is located and where it is going via satellite. The use of satellites equipped with MODIS (Moderate Resolution Imaging Spectroradiometer) Which are used to create high resolution pictures of the area while being able to actually see where the fire is located without interference from weather conditions and smoke [3].

*2.3.2 Meteorological Variables.* These variables are really the main cause of fire spread. Many environmental factors such as temperature and humidity play a decent role in the behavior of the fire. However, the largest contributing factor to fire spread and overall behavior is wind. When wind speeds reduce typically fires are much easier to control. In harnessing data about wind conditions in Southern California, there maybe hope for creating real time, accurate fire spread data. Thus helping fire fighters react appropriately to where the fire is going, rather than just reacting to the fire itself. therefore, helping save money in property damage and live, and resources.

## 3 FIREMAP

Fire map is an interactive web based application in which you can simulate wildfire models. It also utilizes WIFIRE's data sets. It allows users to explore the map and work with multiple layers. The

layers can be changed, or have their opacities changed, allowing for multiple layers to be built upon one another. This give the user customizable data and better insight in the data.

### 3.1 Pylaski

To make Firemap to be interactive and useful to users it must be able to provide all the different types data from the multiple sources where the data is located. To do this Pylaski uses a rest service to query internal as well as external sources for date. It send a query for each individual request and then returns all the data at once back to the client. Pylaski is also able to use different schema's as well as different formats. This works fairly well creating an idea prediction. However as it is collecting massive amounts of data it is unable to query all of the data request because of size limitations. Though limiting it is most likely able to satisfy the requirements of the user [4].

### 3.2 GeoServer

GeoServer is a open source server for geographical datasets. It allows you to get data from a large number of geospatial data collection agencies. With these you can pull up a large number of maps.

**Historical Fires:** These are file that were collected on previous fire parameters. There were recoded for the whole United States from the years 2000-2015. In addition there are many records from California specifically from 1878-2014. All which if clicked on, tell you descriptions of the fire in detail.

**Satellite Detection:** Used thermal imaging and MODIS satellites to produce high resolution photos of fires from above. These fly over about two time a day, and are uploaded to NASA.

**Smoke Areas:** Show smoke plumes and help possibly identify the start of new wildfires. Data comes from NOAA

**Red Flag Areas:** Red flag areas are places where conditions are right for a possible fire occur. These are typically indicated by high wind speed or low relative humidity. this data comes from the National Wildfire Coordinating Group.

**Census Block:** The census block comes from the US Census which keep track of population density. Allowing for consideration of people affected by a fire.

**Surface Fuels:** Used thirteen ways to categorize fuel for fires through vegetation. This is a large factor in how fires will spread as they need easy to burn fuel to keep burning.

**Canopy Cover:** The percentage of forest floor covered by trees. Crutal in analyzing model crown fires which spread from top layers of trees.

**Camera Viewsheds:** Geographical areas with visibility of surrounding areas in which the HPWREN and AlertTahoe cameras are used to visualize the area of the fire.

## 4 WIFIRE

Wildfires in the right setting are beneficial to health of an ecosystem. However, when you put millions of people within that ecosystem, it becomes a problem. This is the problem that the WIFIRE project started at University of California San Diego is trying to remedy.

The have created an end to end cyber infrastructure system to create real time prediction, visualization for wildfires in San Diego Counter. They use real time data from sensors and satellite for observation as well as meteorological study in order to create up to date predictions. Being that San Diego county sees many wildfires, it is a prime location for testing this project out. The most recent fire in 2007 cost over one billion dollars and causes half a million people to be evacuated from their homes[1]. In order to collect the correct and most up to date information for tracking a wild fire they call upon the High Performance Wireless Research and Education Network known as HPWREN. In addition, the use various partner sites that also collect large amounts of data. These come from various types of data, such as, meteorology, vision and audio and hydrology. To make all of relevant and precise for the end users, such as the fire fighters.

### 4.1 Systems

Their systems for collecting this heterogeneous data uses a few different layers of architecture. There is a pipeline composed of many different sensors from the San Diego county, such as ground based sensors, SDG& E (San Diego Gas & Electric), Water Meteorology Stations and the HPWREN as well as collect satellite imagery. With all this heterogeneous data coming in the needed a way to process all the data. They achieved this with an open source workflow system called Kepler which organized the execution of the real time data processed. Then that system distributed it to a portal for end user consumption. This system is meant to handle large data input as well as execute computation in parallel so all data given is up to date [2].

### 4.2 Subsystems

The most important part of all this system is the data pipeline. Which in their workflow is called the Data Communication Subsystem. This system handles the large input of data and works with integration many different data sets and information from imaging, whether it be satellite or camera. The influx of data is come is produced by three types of communication systems. Meteorology stations, still cameras around San Diego County, and Satellite and Aerial data. This data communication layer uses REST services to take in the data to the database and the who process is done in memory to keep processing time as low as possible.

### 4.3 Data Mining

Big data today would not be what it is today if not for the fundamental process of data mining. Predicting trends based out of huge amounts of data, in order to understand exploit that data for practical use and application. WIFIRE also uses this methodology to find data relating to Santa Ana winds. Analyzing the data from the meteorological stations from around the county can help them realize when and where a Santa Ana wind possible. This data would be useful to the firefighter, because it would allow them to know the fire may have a chance to spread further. In addition this data can also be put into fire modeling to help predict wildfires movement even better in the future.

#### 4.4 Visualization

Using visual this project's goal is provide the most transparent data to first responders. Using satellite imagery to understand areas that have been previously scarred by fires and less likely to burn, while identifying areas that may be more susceptible to burning. This uses the topography and analization from previous fires to give them a better picture of how fires moved throughout the county. In addition to using the aerial data, the project is trying to create an application for pedestrians and bystanders to use to help track active fires from the ground. The premise of this idea is that if enough users take photos they can recreate where that photo was taken in a 3D space and track the fire that way. It would also help identify the fires over all perimeter. Once again helping first responders, more effectively fight the fires and know the whole area it covers and where it is heading.

#### 4.5 Assimilation

Most programs on the market are based off of just pure data, and produce a simulation base of that. The problem with that is fires and still not predictable. Using the behavior of real fires from real time data, and compares that against stimulation's, to correct the errors based off of the simulation alone. The system uses recursive algorithms created by Extended Kalman Filter and a Ensemble Kalman Filter. It uses existing fire models such as those in FARSITE in a separate workflow. Allowing for the real time fire date to run in process, and compare them against the simulation data previously provided. This allows them to be compared and move them to the update step, which is where the data assimilation takes place. Correcting previous errors for the simulation, allowing the simulations for future predictions to be much more accurate.

#### 4.6 Santa Ana Winds

Due to the amount of wildfires that spur up per year in the area, the amount of data received over wildfire may very. However, one of the biggest factors in wildfires and their spreading ability is with Santa Ana Winds. They are based off a couple key conditions, such as direction, speed and relative humidity. All of which create extremely dangerous situations for fires. Luckily there are many sensor around San Diego County that can detect this happening. The issues is however, is when an alert is sent, it is only for a specific point. With the kepler workflow they are able to take in data from those individual stations and run it through a program called WindNinja which can take vegetation, topography and weather stations into account produce a mapping of the areas affected. These are ran through the workflow breaking down tiles from the stations in 50km by 50km blocks for marking the wind occurrence. This then runs through a data processing software such as Spark or Hadoop.

### 5 CONCLUSION

In conclusion, the use of big data in fire prevention and tracking is an impressive system. It incorporates a multitude of sensors and data from various sources. Creating predictions of fires and creating data models which will help first responders saves lives, and protect property from damage. This application will allow efficient of firefighter resources and allow for them to be one step ahead of

the fire. Which until now was impossible. Only when the weather changed, were we able to change and react. With the predictive modeling and machine learning, the models will continue to get more accurate as the topography and fire interactions continue into the future. While this system will not prevent all wild fires, it will be instrumental in slowing them down and containing them. Once again man will try to tame fire as we have tried to do since the beginning of our existence.

### ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and the TA's for their help and support.

### REFERENCES

- [1] 2007. Cost of California Wildfires Is More than 1 Billion. (October 2007). <https://www.npr.org/templates/story/story.php?storyId=15603441>
- [2] Ilkay Altintas, Jessica Block, Raymond de Callafon, Daniel Crawl, Charles Cowart, Amarnath Gupta, Mai Nguyen, Hans-Werner Braun, Jurgen Schulze, Michael Gollner, Arnaud Trouve, and Larry Smarr. 2015. Towards an Integrated Cyber-infrastructure for Scalable Data-driven Monitoring, Dynamic Prediction and Resilience of Wildfires. *Procedia Computer Science* 51, Supplement C (2015), 1633 – 1642. <https://doi.org/10.1016/j.procs.2015.05.296> International Conference On Computational Science, ICCS 2015.
- [3] C. Brun, T. Artes, A. Cencerrado, T. Margalef, and A. Corts. 2017. A High Performance Computing Framework for Continental-Scale Forest Fire Spread Prediction. *Procedia Computer Science* 108, Supplement C (2017), 1712 – 1721. <https://doi.org/10.1016/j.procs.2017.05.258> International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.
- [4] Daniel Crawl, Jessica Block, Kai Lin, and Ilkay Altintas. 2017. Firemap: A Dynamic Data-Driven Predictive Wildfire Modeling and Visualization Environment. *Procedia Computer Science* 108, Supplement C (2017), 2230 – 2239. <https://doi.org/10.1016/j.procs.2017.05.174> International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

```
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 NPR
Warning--to sort, need author or key in NPR
Warning--no key, author in NPR
Warning--no key, author in NPR
Warning--no key, author in NPR
Warning--empty author in NPR
(There were 6 warnings)
```

```
bibtext _ label error
```

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

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

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

```
=====
[2017-12-10 13.53.17] 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 1.0s.
```

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

```
=====
name: Zachary Meier
hid: 346
paper1: Nov 25 17 100%
paper2: Nov 28 17 100%
```

```
project: Dec 08 17 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 346 project (null) 2573 report.tex
wc 346 project (null) 2615 report.pdf
wc 346 project (null) 555 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

```
3: \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 in NPR
Warning--to sort, need author or key in NPR
Warning--no key, author in NPR
Warning--no key, author in NPR
Warning--no key, author in NPR
Warning--empty author in NPR
(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
```

# Gerrymandering Detection Using Data Analysis

Kevin Duffy  
Indiana University  
4014 E. Stop 10 Rd.  
Indianapolis, Indiana 46237  
kevduffy@iu.edu

## ABSTRACT

Can the evergreen issue of partisan gerrymandering be solved using data and algorithms? That question is closer to being solved than ever, as a method developed by University of Chicago researchers in 2014 is currently pending approval by the United States Supreme Court. We examine the method, known as the efficiency gap model, using data from Indiana's recent State Senate and House elections. We then evaluate the effectiveness of this model, take stock of any substantial critiques, and suggest courses of further research.

## KEYWORDS

Big Data, Elections, Gerrymandering, Voting, Efficiency gap, i523, HID 310

## 1 INTRODUCTION

In 1964 the Supreme Court established in constitutional law the principle of "one-person, one-vote".[10] The idea appears self-evident; the value of any citizen's vote is equal to that of any other citizen's vote. But could anything still exist in our institutions of our democracy that resists this principle? Beyond obvious impediments to voting such as the since-repealed prohibition on women or racial minorities voting, what other barriers could exist? And why has "one-person, one-vote" become an issue as important, and contentious, as ever?

The barrier, many will argue [17][24][2], lies in the concept of "gerrymandering", or the manipulation of legislative district lines for the benefit of one political player over another. But determining whether something is gerrymandered has proven to be a difficult task. And even once you decide something is gerrymandered, what can be done about it? Answers to these questions may be coming in the form of both advanced data analysis, and simple arithmetic.

In this paper we will lay the foundation of the issue: what is gerrymandering, when has it been used successfully, and what have people done to attempt to curb it? We will then explore possible tests to identify gerrymandering, particularly the efficiency gap method which is currently be litigated before the United States Supreme Court. We will replicate the efficiency gap method using data from Indiana's recent State Senate and House elections to determine the level of gerrymandering in the state's legislative maps, according to the model and suggest avenues for further study. We will evaluate these results, seek to find corroborating evidence for the model's accuracy, and suggest further courses of study.

## 2 WHAT IS GERRYMANDERING?

The building blocks of America's democratic republic are legislative districts. Any one American lives in three overlapping districts: their State Senate district, their State House of Representatives

district, and their U.S. House of Representatives (congressional) district. There is a set limit of 435 congressional districts dispersed to the 50 states based on population measured at the last U.S. Census. Within each state, each congressional district must be exactly the same size. Across the nation as a whole, the average population of a district is 710,000 residents. State House and Senate districts are given a 10 percent population range, as long as it is not systematically used to give one party an advantage. The federal government leaves it to the states to draw the lines all three of these maps. While there is some variance, most states have these lines drawn by the state legislature themselves. The lines are redrawn every ten years, after the census is conducted, in a process called redistricting.[14][15][12]

Gerrymandering, simply put, is the process by which a political party in power uses redistricting to "manipulate district boundaries to create maps that systematically advantage the party in control and lock in an advantage for the party in future elections", according to the NYU Brennan Center of Justice.[23] In other words, the political party seeks to maximize the efficiency of each vote for their candidates, while decreasing the efficiency of the opposite party. This is done primarily through by "cracking" and "packing" the district maps:

- *Cracking* means splitting up a bloc of voters loyal to one party into several other districts, thus diluting the power of their collective vote and maximizing the amount of districts the preferred party is competitive in.
- *Packing* means concentrating a bloc of voters loyal to the opposing party into one district, giving them an overwhelming share of the vote in that district, but decreasing their power in several other districts. [24]

The overall effect of cracking and packing is maximize the "wasted votes" of the disfavored party, while minimizing those of the favored party. The concept of wasted votes is crucial to the model we will examine further. A wasted vote can be defined as one of the following:

- For the winning candidate (party A), a wasted vote is any vote beyond the threshold needed to win, or 50 percent of the vote.
- Every vote for the losing candidate (party B) is considered wasted, as the vote did not net a seat for that party.[24]

You can clearly see the potential for cracking and packing to greatly effect the amount of wasted votes a certain party receives. When a party's voters are cracked into several different districts, the net effect is that their votes go to various losing candidates, thus the votes are all wasted. When a party's voters are packed into one

concentrated district, the net effect is that their votes overwhelmingly elect just one candidate, well beyond the needed threshold to win. Thus, all the votes beyond that threshold are wasted.

## 2.1 History

This is not merely a theoretical exercise. Gerrymandering has been utilized throughout the history of the United States by virtually every political party that has been in power. In fact, the term "gerrymander" dates back to 1812, when the *Boston Gazette* used the phrase to decry a unfairly-drawn redistricting plan signed into law by Massachusetts Governor Elbridge Gerry.[20] A famous political cartoon depicts one particularly contrived looking district as a dragon-esque creature, while others compared its shape to that of a salamander. The colloquial term became "gerry-mander" after the governor who enabled such a result.

The utilization and effectiveness of gerrymandering does not appear to have lessened with time. One of the most effective uses of the practice happened earlier this decade.

In 2008, the national Republican Party was in a dire position. Barack Obama had just been elected president in a sweep that included control of both houses of Congress. They held the House of Representatives by the largest margin seen in almost 20 years.[1] Journalist Michael Grunwald presented a grim narrative for the party in Time magazine in 2009, writing that "polls suggest that only one-fourth of the electorate considers itself Republican, that independents are trending Democratic and that as few as five states have solid Republican pluralities." In addition, he pointed out that the overall population was decreasing in demographics that had proven to be solidly Republican - "less white, less rural, less Christian".[21]

Fast-forward to 2017: Republicans control the White House and Congress, owning the House by almost as large a margin as the Democrats did just 8 years earlier.[1] What happened to produce results that so starkly contrast with the outlook Grunwald predicted?

## 2.2 REDMAP

In the wake of the 2008 election, the Republican State Leadership Committee launched the Redistricting Majority Project (REDMAP).[4] The strategy of the plan was brilliant in its simplicity: use their funding to target state legislature races in order to control as many state legislatures as possible when the next redistricting occurred in 2011. In most states, the task of drawing new district boundaries is left to the state legislatures.[14] Redistricting occurs after each census to reflect changes in population densities and demographics. The strategy paid off in giving Republicans control of the redistricting process in many of the states. From there, partisan politics began its work.

REDMAP was a clear success as evidenced by the ensuing 2012 election, the year in which Barack Obama won reelection. In Michigan, for example, voters cast 240,000 more votes overall for Democrats in congressional races, but 9 Republicans were elected and only 5 Democrats. In Pennsylvania, voters cast 83,000 more votes overall for Democrats in congressional races, but 13 Republicans were elected and only 5 Democrats. Across the nation, Republicans won 54 percent of house seats and 58 out of 99 state legislative chambers, while winning only 8 out of 33 Senate races (which are gerrymander-proof as they do not rely on district lines).[4]

The effect of gerrymandering is evident in the results of REDMAP. Republicans maximized the wasted votes of the Democrats and minimized the wasted votes of the Republicans. Although in many cases Democrats had more votes statewide, more Republicans candidates were sent to Congress.

Gerrymandering is an issue seen by many in both parties as problematic. Representative Brian Fitzpatrick (R-PA) wrote in *The Hill* that gerrymandering has caused the nation to stray from its ideal of representative leadership, as it has "has undermined community-focused representation by forcing lawmakers to ideological extremes and exacerbating electoral complacency that causes lawmakers to focus on accumulating power rather than serving constituents".[19]

However, despite bipartisan efforts in the legislature such as those lauded by Fitzpatrick, the most promising avenue for curbing gerrymandering may lie in a different branch of government: the United States Supreme Court.

## 2.3 The Supreme Court and Gerrymandering

The court has already banned racial gerrymandering in the decision on *Thornburg v. Gingles* in 1986. They determined that a district map in North Carolina violated the Voting Rights Act by gerrymandering the districts in a way that unfairly diluted the power of black voters.[2]

But up until recently, partisan gerrymandering has been left to the states to police themselves. The Supreme Court has heard around 50 cases in its history imploring the court to intervene against a partisan gerrymandered map, and each time has deferred.[24] The last major case was *Vieth v. Jubelirer* in 2004. Justice Antonin Scalia, since deceased, delivered the majority opinion stating that the courts were not responsible for partisan gerrymandered maps as they were "non-justiciable".[17]

Justice Anthony Kennedy is reliable in his unreliability - he serves as the court's "swing vote", as court observers are often unsure which way he'll fall on a given issue until the decision is handed down. Given the court's ideological polarity often leads to close votes, this arguably makes him the most powerful justice on the bench. In *Vieth*, Kennedy voted along with the majority opinion that upheld the allegedly-gerrymandered maps. However, he left open the possibility of the court adjudicating gerrymanders, if a clear standard could be found for determining whether a map is gerrymandered or not.[17]

Such a standard has not been apparent until, perhaps, now.

## 3 THE EFFICIENCY GAP

In 2017, a case reached the Supreme Court alleging that the Wisconsin State Assembly was gerrymandered in such an extremely partisan way as to render it unconstitutional. The plaintiffs, savvy enough to recognize Justice Kennedy as the potential swing vote and remembering his desire for a clear standard, argued their case using the "efficiency gap" method.[17]

The efficiency gap method was developed by University of Chicago law professor Nicholas Stephanopoulos and Eric McGhee, a researcher at the Public Policy Institute of California.[24]

The efficiency gap is a relatively simple formula, based on the aforementioned concept of wasted votes. The formula takes the

total statewide wasted votes of party A and subtracts the total statewide wasted votes of party B, and then divides that number by the total number of statewide votes to find the efficiency gap score.[25]

$$Gap = \frac{Waste(A) - Waste(B)}{TotalVotes}$$

Another way it can be written:

$$\text{Gap} = (\text{Seat margin}) - (2 \times \text{Vote margin})$$

The "seat margin" is the percentage of seats you win from the statewide allotment minus 50 percent, and the "vote margin" is the total percentage of the vote you win minus 50 percent. A negative result means the map is biased against you.[25] This is a helpful format when we begin measuring the net effect of gerrymandering in congressional districts.

If the two parties have similar numbers of wasted votes, or neither party has a significant amount of wasted votes, the efficiency gap score for that state will be low indicating acceptable levels of map bias. However, if one party has a disproportionate number of wasted votes compared to its opponent, the result will be a higher efficiency gap score indicating unacceptable levels of map bias.

This method captures the effects of both cracking and packing: packing will be detected by the wasted votes from an excessive victory, and cracking will be detected excessive amounts of losing votes statewide.

Let's apply this to an example. Let's say Party A and Party B are competing in a state with ten congressional districts of 100 people each.

District	Party A votes	Party B votes
01	<b>90</b>	10
02	49	<b>51</b>
03	45	<b>55</b>
04	<b>95</b>	5
05	45	<b>55</b>
06	<b>90</b>	10
07	49	<b>51</b>
08	45	<b>55</b>
09	<b>95</b>	5
10	45	<b>55</b>

In Districts 01, 04, 06, and 09, Party A wins by an overwhelming margin. In the rest of the districts, Party B wins by narrow margins. This results in more votes being cast for Party A statewide, but Party B gets more seats:

	Party A	Party B
Votes	648	352
Seats	4	6

These races result in an overwhelming amount of wasted votes for Party A, and a minimal amount for Party B.

District	Party A Wasted votes	Party B Wasted votes
01	40	10
02	49	1
03	45	5
04	45	5
05	45	5
01	40	10
02	49	1
03	45	5
04	45	5
05	45	5
<b>Total</b>	<b>448</b>	<b>52</b>

We can see that Party A has many more wasted votes than Party B, indicating the map may be drawn to minimize the efficiency of Party A. We then add up the total number of votes cast statewide and plug these numbers into our efficiency gap formula:

$$\frac{448 - 52}{1000} = \frac{396}{1000} = 0.396$$

So our efficiency gap, written as a percentage, is 39.6 percent. The map is clearly tilted in favor of Party B. But is it considered illegal gerrymandering? In their paper, the authors establish thresholds for when an efficiency gap indicates levels of illegal gerrymandering:

- For state legislature maps, an efficiency gap score above eight percent is considered illegally gerrymandered. The mere percentage is used as each legislature is an entity unto itself, elected wholly by voters in the state. This along with variances in size among state legislatures, makes efficiency gap the best way to normalize disparate state houses for comparison.
- For congressional maps, a state is considered illegally gerrymandered if the map costs a party two seats. In contrast to state houses, the authors contend, "aggregate House seats are the parties' main objective". In that regard, seats are the best way to normalize disparate state sizes for comparison.[26]

If we write our formula in the format  $(\text{Seat margin}) - (2 \times \text{Vote margin})$ , we can measure how many seats were lost as a result of the biased map. In this example, Party A won 64.8 percent of the vote, but was awarded only 4 out of the 10 seats. For Party A:

$$\begin{aligned} (.40-.50) - (2 \times .148) \\ -.10 - .296 \\ -.396 \end{aligned}$$

Now let's give Party A enough seats to make the efficiency gap score as close to 0 as possible. We will say that Party A in this alternate scenario received 8 seats, represented as .80 in the seat share value:

$$\begin{aligned} (.80-.50) - (2 \times .148) \\ -.30 - .296 \\ -.004 \end{aligned}$$

We have brought the score effectively to 0. So using this formula, we have determined that the efficiency gap derived from the biased map cost Party A a total of 4 seats, well above the threshold for illegal gerrymandering.

The question remains whether this standard will be used to measure map bias and judge gerrymandering. During oral arguments

for *Whitford* in October 2017, Chief Justice John Roberts referred to the theory as "sociological gobbledegook".[13] But some court observers are anticipating Justice Kennedy, the swing vote, to vote in favor of the efficiency gap test.[30]

### 3.1 Criticism

As with any politically-charged debate, the efficiency gap has drawn criticism for simplifying the electoral process too much to come to its result. Critics contend that there are several factors the method ignores that may explain the phenomenon of gerrymandering. Two such critics, Chris Winkelman, an attorney for the National Republican Congressional Committee and Phillip Gordon, an outside attorney, filed a brief with the Supreme Court in the *Gill v. Whitford* case exposing what they saw as flaws in the efficiency gap formula.[29] These are indicative of the major arguments against the efficiency gap that we have found.

- (1) The method assumes that voters' party loyalties are static. In predicting the future, the model assumes that voters never change their minds and are not swayed by the contextual candidates involved. Winkelman and Gordon point to the 2012 and 2016 elections. They contend that studies have shown between 11 to 15 percent of voters chose Barack Obama, a Democrat, in 2012 and then chose Donald Trump, a Republican, in 2016. In addition, in 2016 12 Democrats won in districts that elected Trump while 23 Republicans won in districts that elected Clinton.
- (2) The method ignores the effect of partisan geometry. The method assumes that partisan loyalties are spread out evenly among a state, when this is not the case. Typically, Democrats tend to concentrate heavily in urban areas, while Republicans are more thinly spread out among rural and suburban areas. This makes drawing maps in a compact and contiguous manner, which most agree is the ideal way to draw a map as opposed to winding, snake-like districts that are clearly to lump chosen groups of voters together, naturally beneficial to Republicans.

It is beyond the scope of this analysis to objectively prove or disprove points for and against the efficiency gap method. We provide counterpoints to show that there is not perfect consensus on this topic; this objective measure rests in the subjective hands of the Supreme Court.

## 4 APPLICATION

We implemented the efficiency gap method into a python application powered by real-world election data in order to determine whether the district maps for Indiana's House of Representatives and State Senate pass or fail the efficiency gap thresholds.

Indiana was chosen arbitrarily, primarily because it is the home to both the author and institution of this paper. In addition, Indiana's legislature gives us clean and uniform data to work with, as there are an even number of senators and representatives elected, no special elections, and no run-off elections to complicate the data. However, after implementation of the application, it became apparent that it was fortunate Indiana was chosen as the results showcase important teaching moments in understanding the efficiency gap and its applications.

Because we were evaluating state legislatures, we did not have to calculate seats lost, so the results are given in raw efficiency gap score.

### 4.1 Data sourcing and cleaning

For our data, we use the election results from the 2016 Indiana House of Representatives races[8], and the 2014 and 2016 Indiana State Senate races[6][9]. The data was collected from Ballotpedia, an online election and candidate encyclopedia. For context, each member of the House is elected every even year for a two-year term. There are 100 representatives. Each member of the Senate is elected to a four-year term, with elections occurring every two years to elect half of the members. There are 50 senators. Thus, to receive a full sample of the House races, we only needed to collect data from one election year. But for the Senate, we needed two election years in order to collect data for the full senate.

There were two complicating factors with the data that needed to be cleaned before implementation into our model:

- (1) One third of the races in 2016 were uncontested, meaning the winning candidate had no opposition to compare to. In 2014, almost half were uncontested. Depending on the county data recording, these are represented in one of two ways.
  - (a) The votes cast for the winner are displayed, resulting in an election that looks like 20,000 votes were cast for candidate A, and 0 votes were cast for candidate B.
  - (b) No votes are displayed, and the winner is simply displayed as a default.

Both of these taken at face value are problematic for our model. In the first case, plugging these results into our model could overstate how many wasted votes there were for the winning candidate, as the model would think that the winner received 20,000 more votes than they needed to, and the loser received no wasted votes. This is unlikely to occur in reality if the opposing party had fielded a candidate. Even if it is not a close race, the loser would accumulate enough votes to alleviate the amount of wasted votes accrued by the winner and increase the amount of wasted votes for the loser.

In the second case, rather than overstating the wasted votes, they are understated. The race is treated as a draw in terms of wasted votes, when uncontested elections would in reality be a major symptom of an efficiency gap and wasted votes should be accrued.

Clearly, they cannot be ignored. The efficiency gap authors provide guidance on what to do with these races. For state house races, they ran a multi-level model using a fixed effect for incumbency and random effects for year, state, and district. If the district had been contested in its past, the value was derived from other districts in the state during that year along with the same district in other years. If not, they had a random draw of random effects. [27]

The results were a mean Democratic vote share of 66 percent for uncontested Democratic candidates, with 90

percent of values falling between 52 and 83 percent. Democratic vote share for races with uncontested Republicans was placed at 36 percent, with 90 percent of values falling between 22 and 43 percent. The authors do not hold this solution to be the be-all-end-all model for computing vote shares of uncontested candidates, as they "encourage scholars to explore a range of imputation techniques."<sup>[27]</sup>

Our solution was to uncritically use the authors' figures of 34 percent share for Republicans in uncontested Democratic seats, and 36 percent share for Democrats in uncontested Republican seats. For those seats that had no winning vote data available, we took the average population of a district, adjusted for that year's vote turnout, and applied the percentages to that number. For those seats with winning vote data, we simply took half of the winning votes as the loser's share of votes.

If further work to be undergone on this application, we would recommend fine tuning these calculations, particularly if one were to specifically focus on a particular state legislature, as vote shares for a given political party would most likely vary from state to state.

- (2) The other complicating factor for this experiment was the existence of third parties. In the efficiency gap calculations, third party votes are ignored, relying on the two-party vote.<sup>[11]</sup> For most districts, the effect that third parties have is marginal:
  - The United States is a two-party system, mostly due its "winner take all" election rules (where the party with the most votes is the singular winner in a given race, whereas a proportional system would give distribute legislative seats proportionally based on vote share). Third parties therefore have a difficult time gaining any sort of power:
    - The highest vote share of any third party in the 2016 presidential election was 4 percent for the Libertarian Party, the highest share the Libertarian Party had ever received in a presidential election.<sup>[16]</sup>
    - There are no third party members of the Indiana House of Representatives and the Indiana State Senate.<sup>[9][8]</sup>
    - There are no third party members of the U.S. House of Representatives.<sup>[1]</sup>
  - Third parties are varied; there is no one singular third party to claim a stake in the redistricting process. Thus, their voice is diluted by diversification.
  - When we establish that we are operating under a binary party system, third parties make no difference in the efficiency gap formula, as a vote cast for a third party candidate is wasted for Democrats and Republicans equally, thus cancelling itself out.

The solution was straightforward - we simply removed third party votes from our calculations and operated under a two-party vote system.

## 4.2 Implementation

The application made moderate use of the Python Pandas module. We began by importing two dataframes: the 2016 House results, and the 2014 and 2016 Senate results combined into one dataframe. Because the efficiency gap is a simple formula, the values needed are similarly simple. The only values needed were the Republican votes and Democratic votes for each district:

district	dvotes	rvotes
1	15561	7780
2	24820	12786
	...	
99	24820	12786
100	14110	7055

With the data imported, the first step is to calculate the wasted votes for both parties. We have two separate functions to calculate Democratic wasted votes and Republican wasted votes.

```
def dwaste(row):
    if row['dvotes'] > row['rvotes']:
        val = row['dvotes'] - ((row['dvotes'] + row['rvotes']) * .5)
    else:
        val = row['dvotes']
    return val
```

The Republican wasted votes function is identical except the 'dvotes' and 'rvotes' values are switched. This portion of the script goes line by line through the dataframe to calculate the wasted votes for each party per district. This needs to be done row by row as wasted votes cannot be found as an aggregate statewide total, but by looking at each individual district race.

Next, we applied the rwaste() and dwaste() functions to our data frames, and then we can get our statewide totals of wasted votes by party:

```
df['rwaste'] = df.apply(rwaste, axis=1)
rtotal = df['rwaste'].sum()
```

From there, we plugged the statewide wasted votes totals into our efficiency gap formula:

```
((dtotal-rtotal)/((df['rvotes'].sum())+df['dvotes'].sum()))*100
```

Finally, we implemented a simple function that serves as our threshold test. If the formula falls above 8, it triggers an "UNCONSTITUTIONAL GERRYMANDER" response:

```
def eg():
    if final > 8
        print("UNCONSTITUTIONAL GERRYMANDER")
    else:
        print("ACCEPTABLE")
```

Since the data used for this particular experiment is small, the application was able to be executed on a personal computer using an Ubuntu virtualbox.

## 4.3 Results

The application revealed that the House of Representatives, with 2,992,624 votes cast, had 838,675 Democratic wasted votes and 657,637 Republican wasted votes resulting in an efficiency gap score of 6.05 percent in favor of Republicans. This falls under the 8 percent threshold, indicating that if the efficiency gap were to be

adopted as a court standard by the Supreme Court, this map would be ruled constitutional.

On the other hand, the Senate, with 2,107,263 votes cast, has 661,509 Democratic wasted votes and 347,122 Republican wasted votes resulting in an efficiency gap score of 15.58 percent in favor of Republicans. This lands well above the gerrymandering threshold. If this standard were to be adopted by the Supreme Court, there is a decent chance the Senate map would be ruled unconstitutional.

There are a few factors to consider that may be used to explain the discrepancy with the Senate vote:

- If we take at face value that the Senate is twice as gerrymandered as the House, a major reason could be district size. The House has twice as many districts as the Senate over the same land area. The more granular a district is, the more difficult it becomes for a map to be gerrymandered, as you have smaller populations and smaller land areas per district. People of similar ideologies and political leanings tend to group together, so with smaller parameters, it becomes more difficult to group some of these individuals with opposing parties in order to "crack" their vote.
- If we are skeptical, the results could be explained by the fact that half of the Senate data was taken from the 2014 midterm election, while all of the House data was taken from the 2016 presidential election.
  - Midterm elections tend to have lower voter turnout, so the data may not be as accurately reflect the true political landscape of a region. In 2014, Indiana had a voter turnout rate of 28 percent, the lowest in the nation that year[5], compared to 58 percent in 2016.[7]
  - Historically, Republicans tend to have higher turnout in midterm elections. Nationally, Republicans were 20 percent more likely to vote in 2010 and 2014 than Democrats were, according to an analysis by the New York Times' Nate Cohn. [18]

Further analysis would need to be done in order to determine if this explanation suffices. It would, in fact, be in the Republicans best interest to find an alternative explanation other than gerrymandering. The efficiency gap authors, in outlining their proposed court test, allow that states above the threshold could show that the the gap was the result of either the "consistent application of legitimate policies", or "inevitable due to the states' underlying political geography".[26]

However, when compared to other Indiana political measures, the Senate result is not especially surprising. While the Indiana is admittedly a "red state", or a state predominantly partial to Republicans, the Senate seat share appears to be out of sync with other measures. Currently, there are only 9 Democrats to 41 Republicans in the Indiana State Senate[6][9], meaning the Democrats have a 18 percent seat share. Compare this with the last three statewide elections for governor in Indiana:

Year	Republican	Democrat
2008	58.8%	40.0%
2012	49.4%	46.5%
2016	51.3%	45.4%

[3] Meanwhile, the House of Representatives has 30 Democrats to 70 Republicans [8], which, while slightly lower than the gubernatorial vote share, can be explained as a "winner's bonus", or a small surplus of votes for the overall winning party. This is accepted by the efficiency gap authors and political scientists as a common feature in American political systems. They also accept that the USA is not a proportional representation system, where vote share corresponds virtually 1:1 with seat share.[25] Therefore measures like gubernatorial vote share versus legislative seat share cannot be used to prove partisan gerrymandering on its own, however, when drastic enough, it can certainly be used as a symptom to correspond with a more direct objective measure.

## 5 LIMITATIONS

Considering we developed an application for one state out of 50, and we did not develop an application for a congressional analysis, the scope of our analysis is limited.

If further research and applications were to be conducted, especially in the realm of big data, it may prove to be beneficial to scrape data from all 50 states to create a time series of state legislature efficiency gap changes across the entire country. This could be coded to import election results dynamically as they are held. This would be useful in observing changes in efficiency gap scores in response to implementations of redistricting plans, identifying ideal redistricting methods, and generating data for use in speculative algorithm-based redistricting applications.

There is also much big data potential in gerrymandering solutions and redistricting methods, while we have been limited to detection. AI and machine learning applications applied to the currently human-driven redistricting effort could prove to be revolutionary for this aspect of our electoral system.

Finally, it is not our intention to suggest that the efficiency gap is the only or even the best method for detecting gerrymandering. It was selected due to its prominence in the major case facing the Supreme Court that could reshape how we measure gerrymandering and conduct redistricting. There are other methods[22][28] that have been proposed that may be able to be compared with the efficiency gap method.

## 6 CONCLUSION

An objective way to measure gerrymandering? Or, as Chief Justice John Roberts so colorfully put it, "sociological gobbledegook"? That definitive answer to that question lies outside the scope of this analysis, and the relevance of that question lies solely with the United States Supreme Court. They have the power, in coming months, to either make this measure the standard with which to measure all district maps moving forward, or toss it aside and continue this country's long history of ignoring partisan gerrymandering as far as the law is concerned.

We have shown how data can be used to transform the debate around partisan gerrymandering, taking what used to be a heated back and forth based on the arguers' political persuasions and elevating it to a debate on which mathematical standard should be used to measure gerrymandering. Though, some still contend that it cannot be measured objectively whatsoever.

We have demonstrated on a preliminary basis some correlation between an excessive efficiency gap score and a disproportionate vote share in the case of the Indiana State Senate.

At the very least, we have sought to prove that the efficiency gap is indeed easy to calculate based on the parameters specified by the method's creators. For someone such as Chief Justice John Roberts, simpler may be better.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and all TA's for their tireless work in ensuring this class goes smoothly.

## REFERENCES

- [1] [n. d.]. Party Divisions of the House of Representatives. ([n. d.]). <http://history.house.gov/Institution/Party-Divisions/Party-Divisions/>
- [2] 1986. *Thornburg v. Gingles*. United States Supreme Court. <https://www.oyez.org/cases/1985/83-1968>
- [3] 2001-2017. Election Results. (2001-2017).
- [4] 2013. 2012 REDMAP Summary Report. (Jan. 2013). <http://www.redistrictingmajorityproject.com/>
- [5] 2014. 2014 midterm election turnout lowest in 70 years. (10 Nov. 2014). <https://www.pbs.org/newshour/politics/2014-midterm-election-turnout-lowest-in-70-years>
- [6] 2014. Indiana State Senate elections, 2014. (2014). [https://ballotpedia.org/Indiana\\_State\\_Senate\\_elections,\\_2014](https://ballotpedia.org/Indiana_State_Senate_elections,_2014)
- [7] 2016. 2016 Election Turnout and Registration. (8 Nov. 2016). <https://www.in.gov/sos/elections/files/2016.General.Election.Turnout.pdf>
- [8] 2016. Indiana House of Representatives elections, 2016. (2016). [https://ballotpedia.org/Indiana\\_House\\_of\\_Representatives.elections,\\_2016](https://ballotpedia.org/Indiana_House_of_Representatives.elections,_2016)
- [9] 2016. Indiana State Senate elections, 2016. (2016). [https://ballotpedia.org/Indiana.State.Senate.elections,\\_2016](https://ballotpedia.org/Indiana.State.Senate.elections,_2016)
- [10] 2017. Constitution Check: What does "one-person, one-vote" mean now? *Constitution Daily* (2017). <https://constitutioncenter.org/blog/constitution-check-what-does-one-person-one-vote-mean-now>
- [11] 2017. *Gill v. Whitford*. United States Supreme Court. <https://www.oyez.org/cases/2017/16-1161>
- [12] 2017. Members of Congress. (2017). <https://www.govtrack.us/congress/members>
- [13] 2017. Oral Arguments - Gill v. Whitford. (October 2017).
- [14] 2017. Who Draws the Maps? Legislative and Congressional Redistricting. *Brennan Center for Justice: Twenty Years* (2017). <https://www.brennancenter.org/analysis/who-draws-maps-states-redrawing-congressional-and-state-district-lines>
- [15] Micah Altman and Michael McDonald. [n. d.]. Equal Population. *Public Mapping Project* ([n. d.]). <http://www.publicmapping.org/what-is-redistricting/redistricting-criteria-equal-population>
- [16] Tessa Berenson. 2016. Third Parties Faded to the Background in a Shocking Election. (9 Nov. 2016). <http://time.com/4562735/third-parties-election-results-gary-johnson-jill-stein-evan-mcmullin/>
- [17] Barry Burden. 2017. Everything you need to know about the Supreme Court's big gerrymandering case. *The Washington Post* (Oct. 2017). [https://www.washingtonpost.com/news/monkey-cage/wp/2017/10/01/everything-you-need-to-know-about-the-supreme-courts-big-gerrymandering-case/?utm\\_term=.066ef1de20d4](https://www.washingtonpost.com/news/monkey-cage/wp/2017/10/01/everything-you-need-to-know-about-the-supreme-courts-big-gerrymandering-case/?utm_term=.066ef1de20d4)
- [18] Nate Cohn. 2017. Democrats Are Bad at Midterm Turnout. That Seems Ready to Change. *The New York Times* (2017). <https://www.nytimes.com/2017/04/05/upshot/democrats-are-bad-at-midterm-turnout-that-seems-ready-to-change.html>
- [19] Brian Fitzpatrick. 2017. Bipartisan, forward-looking solutions on redistricting. *The Hill* (Sept. 2017). <http://thehill.com/blogs/congress-blog/politics/349704-bipartisan-forward-looking-solutions-on-redistricting>
- [20] Elmer Cummings Griffith. 1907. *The Rise and Development of the Gerrymander*. Scott, Foresman.
- [21] Michael Grunwald. 2009. One Year Ago: The Republicans in Distress. *Time* (May 2009). <http://content.time.com/time/magazine/article/0,9171,1896736,00.html>
- [22] Daniel Z. Levin. 1988. Measuring a Gerrymander. *Michigan Journal of Political Science* 9 (1988), 63–67.
- [23] Laura Royden and Michael Li. 2017. Extreme Maps. *Brennan Center for Justice: Twenty Years* (2017).
- [24] Nicholas Stephanopoulos and Eric McGhee. 2014. Partisan Gerrymandering and the Efficiency Gap. *The University of Chicago Law Review* (2014).
- [25] Nicholas Stephanopoulos and Eric McGhee. 2014. Partisan Gerrymandering and the Efficiency Gap. *The University of Chicago Law Review* (2014), 850.
- [26] Nicholas Stephanopoulos and Eric McGhee. 2014. Partisan Gerrymandering and the Efficiency Gap. *The University of Chicago Law Review* (2014), 885.
- [27] Nicholas Stephanopoulos and Eric McGhee. 2014. Partisan Gerrymandering and the Efficiency Gap. *The University of Chicago Law Review* (2014), 866–867.
- [28] Gregory S. Warrington. 2017. Quantifying gerrymandering using the vote distribution. (15 May 2017). <https://arxiv.org/pdf/1705.09393.pdf>
- [29] Chris Winkelman and Phillip Gordon. 2017. Symposium: Mind the gap? The efficiency gap, its failures and the "problem" of geography and choice in redistricting. *SCOTUSblog* (8 Aug. 2017). <http://www.scotusblog.com/2017/08/symposium-mind-gap-efficiency-gap-failures-problem-geography-choice-redistricting/>
- [30] Richard Wolf. 2017. Analysis: Supreme Court debates politics, and silence speaks volumes. (3 Oct. 2017). <https://www.usatoday.com/story/news/politics/2017/10/03/supreme-court-debates-politics-and-kennedys-silence-speaks-volumes/725185001/>

## A ISSUES

DONE:

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

### A.1 Assignment Submission Issues

DONE:

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

### A.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

### A.3 Formatting

DONE:

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

DONE:

Other formatting issues

### A.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

<p><b>DONE:</b> Do not use the word <i>I</i> instead use <i>we</i> even if you are the sole author</p> <p><b>DONE:</b> Do not use the phrase <i>In this paper/report we show</i> instead use <i>We show</i>. It is not important if this is a paper or a report and does not need to be mentioned</p> <p><b>DONE:</b> If you want to say <i>and</i> do not use &amp; but use the word <i>and</i></p> <p><b>DONE:</b> Use a space after . , :</p> <p><b>DONE:</b> When using a section command, the section title is not written in all-caps as format does this for you</p>	<p><b>DONE:</b> Incorrect README file</p> <p><b>DONE:</b> 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</p> <p><b>DONE:</b> The paper has less than 2 pages of text, i.e. excluding images, tables and figures</p> <p><b>DONE:</b> The paper has more than 6 pages of text, i.e. excluding images, tables and figures</p> <p><b>DONE:</b> Do not artificially inflate your paper if you are below the page limit</p>
---	--

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

## A.5 Citation Issues and Plagiarism

<p><b>DONE:</b> It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class</p> <p><b>DONE:</b> Claims made without citations provided</p> <p><b>DONE:</b> Need to paraphrase long quotations (whole sentences or longer)</p> <p><b>DONE:</b> Need to quote directly cited material</p>	<p><b>A.8 Details about the Figures and Tables</b></p> <p><b>DONE:</b> Capitalization errors in referring to captions, e.g. Figure 1, Table 2</p> <p><b>DONE:</b> Do use <i>label</i> and <i>ref</i> to automatically create figure numbers</p> <p><b>DONE:</b> Wrong placement of figure caption. They should be on the bottom of the figure</p> <p><b>DONE:</b> Wrong placement of table caption. They should be on the top of the table</p> <p><b>DONE:</b> Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, jpg</p> <p><b>DONE:</b> Do not submit eps images. Instead, convert them to PDF</p> <p><b>DONE:</b> The image files must be in a single directory named "images"</p> <p><b>DONE:</b> In case there is a powerpoint in the submission, the image must be exported as PDF</p> <p><b>DONE:</b> Make the figures large enough so we can read the details. If needed make the figure over two columns</p> <p><b>DONE:</b> 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.</p>
---	--

## A.6 Character Errors

<p><b>DONE:</b> Erroneous use of quotation marks, i.e. use "quotes", instead of "</p> <p><b>DONE:</b> To emphasize a word, use <i>emphasize</i> and not "quote"</p> <p><b>DONE:</b> When using the characters &amp; # % - put a backslash before them so that they show up correctly</p> <p><b>DONE:</b> 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.</p> <p><b>DONE:</b> 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</p>
--

## A.7 Structural Issues

<p><b>DONE:</b> Acknowledgement section missing</p>
---

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
```

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 232 of file report.bib  
:  
: url = {http://www.levin.rutgers.edu/research/gerrymandering-paper.pdf}  
(Error may have been on previous line)  
I'm skipping whatever remains of this entry  
I was expecting a ',' or a '}'---line 248 of file report.bib  
:  
: publisher = {Indiana Secretary of State}  
(Error may have been on previous line)  
I'm skipping whatever remains of this entry  
Warning--no key, author, or editor in oneperson  
Warning--no author, editor, organization, or key in oneperson  
Warning--to sort, need author, editor, or key in oneperson  
Warning--no key, author, or editor in thornburg  
Warning--no author, editor, organization, or key in thornburg  
Warning--to sort, need author, editor, or key in thornburg  
Warning--no key, author, or editor in maps  
Warning--no author, editor, organization, or key in maps  
Warning--to sort, need author, editor, or key in maps  
Warning--no key, author in govtrack  
Warning--no author, editor, organization, or key in govtrack  
Warning--to sort, need author or key in govtrack  
Warning--no key, author in house  
Warning--no author, editor, organization, or key in house  
Warning--to sort, need author or key in house  
Warning--no key, author in redmap  
Warning--no author, editor, organization, or key in redmap  
Warning--to sort, need author or key in redmap  
Warning--no key, author, or editor in gilltranscript  
Warning--no author, editor, organization, or key in gilltranscript  
Warning--to sort, need author, editor, or key in gilltranscript  
Warning--no key, author in houserelations  
Warning--no author, editor, organization, or key in houserelations  
Warning--to sort, need author or key in houserelations  
Warning--no key, author in senatorresults2014  
Warning--no author, editor, organization, or key in senatorresults2014  
Warning--to sort, need author or key in senatorresults2014  
Warning--no key, author in senatorresults2016

Warning--no author, editor, organization, or key in senateresults2016  
Warning--to sort, need author or key in senateresults2016  
Warning--no key, author, or editor in gill  
Warning--no author, editor, organization, or key in gill  
Warning--to sort, need author, editor, or key in gill  
Warning--no key, author in 2014turnout  
Warning--no author, editor, organization, or key in 2014turnout  
Warning--to sort, need author or key in 2014turnout  
Warning--no key, author in 2016turnout  
Warning--no author, editor, organization, or key in 2016turnout  
Warning--to sort, need author or key in 2016turnout  
Warning--no key, author in sos  
Warning--no author, editor, organization, or key in sos  
Warning--to sort, need author or key in sos  
Warning--no key, author in 2014turnout  
Warning--no key, author in 2014turnout  
Warning--no key, author in 2016turnout  
Warning--no key, author in 2016turnout  
Warning--no key, author in 2016turnout  
Warning--no key, author, or editor in gill  
Warning--no key, author, or editor in gill  
Warning--no key, author, or editor in gilltranscript  
Warning--no key, author in govtrack  
Warning--no key, author in govtrack  
Warning--no key, author in houserelts  
Warning--no key, author in houserelts  
Warning--no key, author in house  
Warning--no key, author in house  
Warning--no key, author, or editor in maps  
Warning--no key, author, or editor in maps  
Warning--no key, author, or editor in oneperson  
Warning--no key, author, or editor in oneperson  
Warning--no key, author in redmap  
Warning--no key, author in redmap  
Warning--no key, author in senateresults2014  
Warning--no key, author in senateresults2014  
Warning--no key, author in senateresults2016  
Warning--no key, author in senateresults2016  
Warning--no key, author in sos  
Warning--no key, author in sos  
Warning--no key, author, or editor in thornburg  
Warning--no key, author, or editor in thornburg  
Warning--no key, author in house  
Warning--no author, editor, organization, or key in house  
Warning--empty author in house  
Warning--empty year in house  
Warning--no key, author, or editor in thornburg

Warning--no author, editor, organization, or key in thornburg  
Warning--empty author and editor in thornburg  
Warning--empty address in thornburg  
Warning--no key, author in sos  
Warning--no author, editor, organization, or key in sos  
Warning--empty author in sos  
Warning--no key, author in redmap  
Warning--no author, editor, organization, or key in redmap  
Warning--empty author in redmap  
Warning--no key, author in 2014turnout  
Warning--no author, editor, organization, or key in 2014turnout  
Warning--empty author in 2014turnout  
Warning--no key, author in senateresults2014  
Warning--no author, editor, organization, or key in senateresults2014  
Warning--empty author in senateresults2014  
Warning--no key, author in 2016turnout  
Warning--no author, editor, organization, or key in 2016turnout  
Warning--empty author in 2016turnout  
Warning--no key, author in houserelts  
Warning--no author, editor, organization, or key in houserelts  
Warning--empty author in houserelts  
Warning--no key, author in senateresults2016  
Warning--no author, editor, organization, or key in senateresults2016  
Warning--empty author in senateresults2016  
Warning--no key, author, or editor in oneperson  
Warning--no author, editor, organization, or key in oneperson  
Warning--neither author and editor supplied for oneperson  
Warning--no number and no volume in oneperson  
Warning--page numbers missing in both pages and numpages fields in oneperson  
Warning--no key, author, or editor in gill  
Warning--no author, editor, organization, or key in gill  
Warning--empty author and editor in gill  
Warning--empty address in gill  
Warning--no key, author in govtrack  
Warning--no author, editor, organization, or key in govtrack  
Warning--empty author in govtrack  
Warning--no key, author, or editor in gilltranscript  
Warning--no author, editor, organization, or key in gilltranscript  
Warning--neither author and editor supplied for gilltranscript  
Warning--no journal in gilltranscript  
Warning--no number and no volume in gilltranscript  
Warning--page numbers missing in both pages and numpages fields in gilltranscript  
Warning--no key, author, or editor in maps  
Warning--no author, editor, organization, or key in maps  
Warning--neither author and editor supplied for maps  
Warning--no number and no volume in maps

```
Warning--page numbers missing in both pages and numpages fields in maps
Warning--empty year in population
Warning--no number and no volume in population
Warning--page numbers missing in both pages and numpages fields in population
Warning--no journal in libertarian
Warning--no number and no volume in libertarian
Warning--page numbers missing in both pages and numpages fields in libertarian
Warning--no number and no volume in wapo
Warning--page numbers missing in both pages and numpages fields in wapo
Warning--no number and no volume in cohn
Warning--page numbers missing in both pages and numpages fields in cohn
Warning--no number and no volume in bipartisan
Warning--page numbers missing in both pages and numpages fields in bipartisan
Warning--empty address in griffith
Warning--no number and no volume in distress
Warning--page numbers missing in both pages and numpages fields in distress
Warning--no number and no volume in brennan
Warning--page numbers missing in both pages and numpages fields in brennan
Warning--no number and no volume in chicago
Warning--page numbers missing in both pages and numpages fields in chicago
Warning--no number and no volume in chicagoformula
Warning--no number and no volume in chicagothreshold
Warning--no number and no volume in chicagouncontested
Warning--no journal in warrington
Warning--no number and no volume in warrington
Warning--page numbers missing in both pages and numpages fields in warrington
Warning--no number and no volume in winkelman
Warning--page numbers missing in both pages and numpages fields in winkelman
Warning--no journal in analysis
Warning--no number and no volume in analysis
Warning--page numbers missing in both pages and numpages fields in analysis
(There were 2 error messages)
make[2]: *** [bibtex] Error 2
```

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

```
[2017-12-10 13.50.43] 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/nxlmf/m/it' instead for sy
Missing character: ""
```

Some font shapes were not available, defaults substituted.

Typesetting of "report.tex" completed in 1.3s.

./README.yml

7:14 warning truthy value is not quoted (truthy)

22:1 error trailing spaces (trailing-spaces)

43:1 error trailing spaces (trailing-spaces)

---

## Compliance Report

---

name: Kevin Duffy

hid: 310

paper1: Nov 20 17 100%

paper2: Nov 06 17 100%

project: Dec 04 17 100%

yamlcheck

---

wordcount

---

9

wc 310 project 9 5705 report.tex

wc 310 project 9 6384 report.pdf

wc 310 project 9 826 report.bib

find "

---

32: In 1964 the Supreme Court established in constitutional law the principle of "one-person, one-vote".\cite{oneperson} The idea appears self-evident; the value of any citizen's vote is equal to that of any other citizen's vote. But could anything still exist in our institutions of our democracy that resists this principle? Beyond obvious impediments to voting such as the since-repealed prohibition on women or racial minorities voting, what other barriers could exist? And why has "one-person, one-vote" become an issue as important, and contentious, as ever?

34: The barrier, many will argue \cite{wapo}\cite{chicago}\cite{thornburg}, lies in the concept of

"gerrymandering", or the manipulation of legislative district lines for the benefit of one political player over another. But determining whether something is gerrymandered has proven to be a difficult task. And even once you decide something is gerrymandered, what can be done about it? Answers to these questions may be coming in the form of both advanced data analysis, and simple arithmetic.

- 41: Gerrymandering, simply put, is the process by which a political party in power uses redistricting to "manipulate district boundaries to create maps that systematically advantage the party in control and lock in an advantage for the party in future elections", according to the NYU Brennan Center of Justice.\cite{brennan} In other words, the political party seeks to maximize the efficiency of each vote for their candidates, while decreasing the efficiency of the opposite party. This is done primarily through by "cracking" and "packing" the district maps:
- 47: The overall effect of cracking and packing is maximize the "wasted votes" of the disfavored party, while minimizing those of the favored party. The concept of wasted votes is crucial to the model we will examine further. A wasted vote can be defined as one of the following:
- 55: This is not merely a theoretical exercise. Gerrymandering has been utilized throughout the history of the United States by virtually every political party that has been in power. In fact, the term "gerrymander" dates back to 1812, when the \textit{Boston Gazette} used the phrase to decry a unfairly-drawn redistricting plan signed into law by Massachusetts Governor Elbridge Gerry.\cite{griffith} A famous political cartoon depicts one particularly contrived looking district as a dragon-esque creature, while others compared its shape to that of a salamander. The colloquial term became "gerry-mander" after the governor who enabled such a result.
- 59: In 2008, the national Republican Party was in a dire position. Barack Obama had just been elected president in a sweep that included control of both houses of Congress. They held the House of Representatives by the largest margin seen in almost 20 years.\cite{house} Journalist Michael Grunwald presented a grim narrative for the party in Time magazine in 2009, writing that "polls suggest that only one-fourth of the electorate considers itself Republican, that independents are trending Democratic and that as few as five states have solid Republican pluralities." In

addition, he pointed out that the overall population was decreasing in demographics that had proven to be solidly Republican - "less white, less rural, less Christian".\cite{distress}

- 71: Gerrymandering is an issue seen by many in both parties as problematic. Representative Brian Fitzpatrick (R-PA) wrote in \textit{The Hill} that gerrymandering has caused the nation to stray from its ideal of representative leadership, as it has "has undermined community-focused representation by forcing lawmakers to ideological extremes and exacerbating electoral complacency that causes lawmakers to focus on accumulating power rather than serving constituents."\cite{bipartisan}
- 78: But up until recently, partisan gerrymandering has been left to the states to police themselves. The Supreme Court has heard around 50 cases in its history imploring the court to intervene against a partisan gerrymandered map, and each time has deferred.\cite{chicago} The last major case was \textit{Vieth v. Jubelirer} in 2004. Justice Antonin Scalia, since deceased, delivered the majority opinion stating that the courts were not responsible for partisan gerrymandered maps as they were "non-justiciable".\cite{wapo}
- 80: Justice Anthony Kennedy is reliable in his unreliability - he serves as the court's "swing vote", as court observers are often unsure which way he'll fall on a given issue until the decision is handed down. Given the court's ideological polarity often leads to close votes, this arguably makes him the most powerful justice on the bench. In \textit{Vieth}, Kennedy voted along with the majority opinion that upheld the allegedly-gerrymandered maps. However, he left open the possibility of the court adjudicating gerrymanders, if a clear standard could be found for determining whether a map is gerrymandered or not.\cite{wapo}
- 85: In 2017, a case reached the Supreme Court alleging that the Wisconsin State Assembly was gerrymandered in such an extremely partisan way as to render it unconstitutional. The plaintiffs, savvy enough to recognize Justice Kennedy as the potential swing vote and remembering his desire for a clear standard, argued their case using the "efficiency gap" method.\cite{wapo}
- 99: The "seat margin" is the percentage of seats you win from the statewide allotment minus 50 percent, and the "vote margin" is the total percentage of the vote you win minus 50 percent. A negative result means the map is biased against you.\cite{chicagoformula}

This is a helpful format when we begin measuring the net effect of gerrymandering in congressional districts.

- 192: \item For congressional maps, a state is considered illegally gerrymandered if the map costs a party two seats. In contrast to state houses, the authors contend, "aggregate House seats are the parties' main objective". In that regard, seats are the best way to normalize disparate state sizes for comparison.\cite{chicagothreshold}
- 217: The question remains whether this standard will be used to measure map bias and judge gerrymandering. During oral arguments for \textit{Whitford} in October 2017, Chief Justice John Roberts referred to the theory as "sociological gobbledegook".\cite{gilltranscript} But some court observers are anticipating Justice Kennedy, the swing vote, to vote in favor of the efficiency gap test.\cite{analysis}
- 253: The results were a mean Democratic vote share of 66 percent for uncontested Democratic candidates, with 90 percent of values falling between 52 and 83 percent. Democratic vote share for races with uncontested Republicans was placed at 36 percent, with 90 percent of values falling between 22 and 43 percent. The authors do not hold this solution to be the be-all-end-all model for computing vote shares of uncontested candidates, as they "encourage scholars to explore a range of imputation techniques."\cite{chicagouncontested}
- 260: \item The United States is a two-party system, mostly due its "winner take all" election rules (where the party with the most votes is the singular winner in a given race, whereas a proportional system would give distribute legislative seats proportionally based on vote share). Third parties therefore have a difficult time gaining any sort of power:
- 324: Finally, we implemented a simple function that serves as our threshold test. If the formula falls above 8, it triggers an "UNCONSTITUTIONAL GERRYMANDER" response:
- 331: print("UNCONSTITUTIONAL GERRYMANDER")
- 335: print("ACCEPTABLE")
- 348: \item If we take at face value that the Senate is twice as gerrymandered as the House, a major reason could be district size. The House has twice as many districts as the Senate over

the same land area. The more granular a district is, the more difficult it becomes for a map to be gerrymandered, as you have smaller populations and smaller land areas per district. People of similar ideologies and political leanings tend to group together, so with smaller parameters, it becomes more difficult to group some of these individuals with opposing parties in order to "crack" their vote.

- 354: Further analysis would need to be done in order to determine if this explanation suffices. It would, in fact, be in the Republicans best interest to find an alternative explanation other than gerrymandering. The efficiency gap authors, in outlining their proposed court test, allow that states above the threshold could show that the the gap was the result of either the "consistent application of legitimate policies", or "inevitable due to the states' underlying political geography."`\cite{chicagothreshold}`
- 357: However, when compared to other Indiana political measures, the Senate result is not especially surprising. While the Indiana is admittedly a "red state", or a state predominantly partial to Republicans, the Senate seat share appears to be out of sync with other measures. Currently, there are only 9 Democrats to 41 Republicans in the Indiana State Senate`\cite{senateresults2014}\cite{senateresults2016}`, meaning the Democrats have a 18 percent seat share. Compare this with the last three statewide elections for governor in Indiana:
- 370: Meanwhile, the House of Representatives has 30 Democrats to 70 Republicans `\cite{houseresults}`, which, while slightly lower than the gubernatorial vote share, can be explained as a "winner's bonus", or a small surplus of votes for the overall winning party. This is accepted by the efficiency gap authors and political scientists as a common feature in American political systems. They also accept that the USA is not a proportional representation system, where vote share corresponds virtually 1:1 with seat share.`\cite{chicagoformula}` Therefore measures like gubernatorial vote share versus legislative seat share cannot be used to prove partisan gerrymandering on its own, however, when drastic enough, it can certainly be used as a symptom to correspond with a more direct objective measure.
- 382: An objective way to measure gerrymandering? Or, as Chief Justice John Roberts so colorfully put it, "sociological gobbledegook"? That definitive answer to that question lies outside the scope of this analysis, and the relevance of that question lies solely

with the United States Supreme Court. They have the power, in coming months, to either make this measure the standard with which to measure all district maps moving forward, or toss it aside and continue this country's long history of ignoring partisan gerrymandering as far as the law is concerned.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

4: \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

---

WARNING: table and above may be used improperly

354: Further analysis would need to be done in order to determine if this explanation suffices. It would, in fact, be in the Republicans best interest to find an alternative explanation other than gerrymandering. The efficiency gap authors, in outlining their proposed court test, allow that states above the threshold could show that the the gap was the result of either the "consistent application of legitimate policies", or "inevitable due to the states' underlying political geography." \cite{chicagothreshold}

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 232 of file report.bib  
:  
: url = {http://www.levin.rutgers.edu/research/gerrymandering-paper.pdf}  
(Error may have been on previous line)  
I'm skipping whatever remains of this entry  
I was expecting a ',' or a '}'---line 248 of file report.bib  
:  
: publisher = {Indiana Secretary of State}  
(Error may have been on previous line)

I'm skipping whatever remains of this entry

Warning--no key, author, or editor in oneperson

Warning--no author, editor, organization, or key in oneperson

Warning--to sort, need author, editor, or key in oneperson

Warning--no key, author, or editor in thornburg

Warning--no author, editor, organization, or key in thornburg

Warning--to sort, need author, editor, or key in thornburg

Warning--no key, author, or editor in maps

Warning--no author, editor, organization, or key in maps

Warning--to sort, need author, editor, or key in maps

Warning--no key, author in govtrack

Warning--no author, editor, organization, or key in govtrack

Warning--to sort, need author or key in govtrack

Warning--no key, author in house

Warning--no author, editor, organization, or key in house

Warning--to sort, need author or key in house

Warning--no key, author in redmap

Warning--no author, editor, organization, or key in redmap

Warning--to sort, need author or key in redmap

Warning--no key, author, or editor in gilltranscript

Warning--no author, editor, organization, or key in gilltranscript

Warning--to sort, need author, editor, or key in gilltranscript

Warning--no key, author in houserresults

Warning--no author, editor, organization, or key in houserresults

Warning--to sort, need author or key in houserresults

Warning--no key, author in senateresults2014

Warning--no author, editor, organization, or key in senateresults2014

Warning--to sort, need author or key in senateresults2014

Warning--no key, author in senateresults2016

Warning--no author, editor, organization, or key in senateresults2016

Warning--to sort, need author or key in senateresults2016

Warning--no key, author, or editor in gill

Warning--no author, editor, organization, or key in gill

Warning--to sort, need author, editor, or key in gill

Warning--no key, author in 2014turnout

Warning--no author, editor, organization, or key in 2014turnout

Warning--to sort, need author or key in 2014turnout

Warning--no key, author in 2016turnout

Warning--no author, editor, organization, or key in 2016turnout

Warning--to sort, need author or key in 2016turnout

Warning--no key, author in sos

Warning--no author, editor, organization, or key in sos

Warning--to sort, need author or key in sos

Warning--no key, author in 2014turnout

Warning--no key, author in 2014turnout

Warning--no key, author in 2016turnout

Warning--no key, author in 2016turnout  
Warning--no key, author, or editor in gill  
Warning--no key, author, or editor in gill  
Warning--no key, author, or editor in gilltranscript  
Warning--no key, author in govtrack  
Warning--no key, author in govtrack  
Warning--no key, author in houserelts  
Warning--no key, author in houserelts  
Warning--no key, author in house  
Warning--no key, author in house  
Warning--no key, author, or editor in maps  
Warning--no key, author, or editor in maps  
Warning--no key, author, or editor in oneperson  
Warning--no key, author, or editor in oneperson  
Warning--no key, author in redmap  
Warning--no key, author in redmap  
Warning--no key, author in senateresults2014  
Warning--no key, author in senateresults2014  
Warning--no key, author in senateresults2016  
Warning--no key, author in senateresults2016  
Warning--no key, author in sos  
Warning--no key, author in sos  
Warning--no key, author, or editor in thornburg  
Warning--no key, author, or editor in thornburg  
Warning--no key, author in house  
Warning--no author, editor, organization, or key in house  
Warning--empty author in house  
Warning--empty year in house  
Warning--no key, author, or editor in thornburg  
Warning--no author, editor, organization, or key in thornburg  
Warning--empty author and editor in thornburg  
Warning--empty address in thornburg  
Warning--no key, author in sos  
Warning--no author, editor, organization, or key in sos  
Warning--empty author in sos  
Warning--no key, author in redmap  
Warning--no author, editor, organization, or key in redmap  
Warning--empty author in redmap  
Warning--no key, author in 2014turnout  
Warning--no author, editor, organization, or key in 2014turnout  
Warning--empty author in 2014turnout  
Warning--no key, author in senateresults2014  
Warning--no author, editor, organization, or key in senateresults2014  
Warning--empty author in senateresults2014  
Warning--no key, author in 2016turnout  
Warning--no author, editor, organization, or key in 2016turnout

Warning--empty author in 2016turnout  
Warning--no key, author in houseresults  
Warning--no author, editor, organization, or key in houseresults  
Warning--empty author in houseresults  
Warning--no key, author in senateresults2016  
Warning--no author, editor, organization, or key in senateresults2016  
Warning--empty author in senateresults2016  
Warning--no key, author, or editor in oneperson  
Warning--no author, editor, organization, or key in oneperson  
Warning--neither author and editor supplied for oneperson  
Warning--no number and no volume in oneperson  
Warning--page numbers missing in both pages and numpages fields in oneperson  
Warning--no key, author, or editor in gill  
Warning--no author, editor, organization, or key in gill  
Warning--empty author and editor in gill  
Warning--empty address in gill  
Warning--no key, author in govtrack  
Warning--no author, editor, organization, or key in govtrack  
Warning--empty author in govtrack  
Warning--no key, author, or editor in gilltranscript  
Warning--no author, editor, organization, or key in gilltranscript  
Warning--neither author and editor supplied for gilltranscript  
Warning--no journal in gilltranscript  
Warning--no number and no volume in gilltranscript  
Warning--page numbers missing in both pages and numpages fields in gilltranscript  
Warning--no key, author, or editor in maps  
Warning--no author, editor, organization, or key in maps  
Warning--neither author and editor supplied for maps  
Warning--no number and no volume in maps  
Warning--page numbers missing in both pages and numpages fields in maps  
Warning--empty year in population  
Warning--no number and no volume in population  
Warning--page numbers missing in both pages and numpages fields in population  
Warning--no journal in libertarian  
Warning--no number and no volume in libertarian  
Warning--page numbers missing in both pages and numpages fields in libertarian  
Warning--no number and no volume in wapo  
Warning--page numbers missing in both pages and numpages fields in wapo  
Warning--no number and no volume in cohn  
Warning--page numbers missing in both pages and numpages fields in cohn  
Warning--no number and no volume in bipartisan  
Warning--page numbers missing in both pages and numpages fields in bipartisan  
Warning--empty address in griffith  
Warning--no number and no volume in distress  
Warning--page numbers missing in both pages and numpages fields in distress  
Warning--no number and no volume in brennan

```
Warning--page numbers missing in both pages and numpages fields in brennan
Warning--no number and no volume in chicago
Warning--page numbers missing in both pages and numpages fields in chicago
Warning--no number and no volume in chicagoformula
Warning--no number and no volume in chicagothreshold
Warning--no number and no volume in chicagouncontested
Warning--no journal in warrington
Warning--no number and no volume in warrington
Warning--page numbers missing in both pages and numpages fields in warrington
Warning--no number and no volume in winkelman
Warning--page numbers missing in both pages and numpages fields in winkelman
Warning--no journal in analysis
Warning--no number and no volume in analysis
Warning--page numbers missing in both pages and numpages fields in analysis
(There were 2 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

# Big data and hearing disability

Rahul Velayutham  
Indiana University Bloomington  
2661 H 7th St  
Bloomington, Indiana 47408  
rahuvela@umail.iu.edu

## ABSTRACT

Big Data is rapidly becoming a crucial component in the majority of the fields, be it from medicine to software. Big data technologies help in processing humongous amounts of data in a rapid manner while enabling us to achieve results fast and accurately. Big data is becoming a key player in the restoration of ecological assets like water, forests and the likes. Real time analysis of assets all over the world and the changes are documented and stored how this data can be used and for what purpose is the penultimate question. We dissect the various stages of the rainwater harvesting process and show how the application of big data to each stage can enhance the process.

## KEYWORDS

Big Data, i523 , HID 232 , Rain Water Harvesting

## 1 INTRODUCTION

Hearing loss, also known as hearing impairment, is a partial or total inability to hear. A deaf person has little to no hearing. Hearing loss may occur in one or both ears. Hearing loss can be temporary or permanent. Hearing loss may be caused by a number of factors, including genetics, ageing, exposure to noise, some infections, birth complications, trauma to the ear, and certain medications or toxins. A common condition that results in hearing loss is chronic ear infections. Certain infections during pregnancy such as syphilis and rubella may also cause hearing loss. Hearing loss is diagnosed when hearing testing finds that a person is unable to hear 25 decibels in at least one ear. Hearing loss can be categorized as mild, moderate, moderate-severe, severe, or profound[7].

To elaborate on the previous paragraph hearing loss can be categorized into two sections, Congenital Hearing Loss and Acquired Hearing Loss. Under Congenital Hearing Loss two chief factors are Genetic and Prenatal Issues. Under Acquired Hearing Loss the chief factors can be listed as Chronic ear infections( also called Otitis Media), Ototoxic drugs (medications that can affect aspects of hearing), Diseases that affect hearing (otosclerosis, Mnire's Disease, meningitis, mumps, etc.), Head injury, Perforated eardrum[1].

More than 50 percent of the time it is believed that genetic factors cause pediatric hearing loss. Genetic or hereditary hearing loss occurs when a gene from one or both of the parents impacts the development of the intricate process of hearing. Genetic issues can affect any portion of the outer, middle or inner ear, and can cause varying degrees of loss. Options for genetic forms of hearing loss vary widely and can range from hearing aids, medication, surgery, cochlear implants or no treatment at all. Prenatal Issues are non-genetic factors that can potentially cause hearing loss before the birth of the child. Factors such as in utero infection, illnesses, toxins

consumed by the mother during pregnancy or cytomegalovirus (CMV) can be passed on to a child in utero and may cause hearing loss. During the birthing process, procedures performed to save a baby's life in an emergency, such as a ventilator or a strong antibiotic, can also affect hearing[7].

As of 2013 hearing loss affects about 1.1 billion people to some degree[7]. It causes disability in 5% (360 to 538 million) and moderate to severe disability in 124 million people [7] . Of those with moderate to severe disability 108 million live in low and middle-income countries. Of those with hearing loss, it began in 65 million during childhood. Those who use sign language and are members of Deaf culture see themselves as having a difference rather than an illness. Most members of Deaf culture oppose attempts to cure deafness and some within this community view cochlear implants with concern as they have the potential to eliminate their culture. The term hearing impairment is often viewed negatively as it emphasizes what people cannot do. Despite all of the solutions and rationalizations being made, it cannot be denied however that hearing loss is becoming an important problem in today's society and one whose numbers is constantly increasing[7].

Big data is perhaps the most interesting technological advancement made in the current era, it has roots in almost all fields right from health care to education to even government policies. It is the far reach that makes big data important, it allows users and clients to make better-informed decisions by taking into account almost all factors. Doctors are looking towards big data to make more accurate diagnostics and look for new medicines, economists are looking towards big data to make more accurate models. The paper will look into how it can enhance some of the solutions provided for those hard of hearing like hearing aids, closed caption etc. It will also suggest enhancements towards preemptively preventing situations that could lead to hearing loss[7].

## 2 BIG DATA IN HEARING AIDS

### 2.1 introduction

Hearing aids are small electronic devices that you wear in or behind your ear. It makes some sounds louder so that a person with hearing loss can listen, communicate, and participate more fully in daily activities. A hearing aid can help people hear more in both quiet and noisy situations. A hearing aid has three basic parts: a microphone, amplifier, and speaker. The microphone receives sound, which converts it into electrical signals and sends them to an amplifier. The amplifier increases the power of the signals and then sends them to the ear through a speaker. [image-hearing aid breakdown] They improve the hearing and speech comprehension of people who have hearing loss that results from damage to the small sensory cells in the inner ear, called hair cells(sensorineural hearing loss). The

damage can occur as a result of disease, aging, or injury from noise or certain medicines. A hearing aid magnifies sound vibrations entering the ear. Surviving hair cells detect the larger vibrations and convert them into neural signals that are passed along to the brain. The greater the damage to a person's hair cells, the more severe the hearing loss, and the greater the hearing aid amplification needed to make up the difference. However, there are practical limits to the amount of amplification a hearing aid can provide. However, if the inner ear is too damaged, a hearing aid would be ineffective.

There are 3 different styles of hearing aids:

a.Behind-the-ear (BTE): Hearing aids consist of a hard plastic case worn behind the ear and connected to a plastic earmold that fits inside the outer ear. The electronic parts are held in the case behind the ear. Sound travels from the hearing aid through the earmold and into the ear. BTE aids are used by people of all ages for mild to profound hearing loss.

b.In-the-ear (ITE): Hearing aids fit completely inside the outer ear and are used for mild to severe hearing loss. The case holding the electronic components is made of hard plastic.

c.Canal: Aids fit into the ear canal and are available in two styles. The in-the-canal (ITC) hearing aid is made to fit the size and shape of a person's ear canal. A completely-in-canal (CIC) hearing aid is nearly hidden in the ear canal. Both types are used for mild to moderately severe hearing loss.

[figure of different hearing aids]

Hearing aids work differently depending on the electronics used. The two main types of electronics are analog and digital.

Analog aids convert sound waves into electrical signals, which are amplified. The aid is programmed by the manufacturer according to the specifications recommended by your audiologist. An audiologist can program the aid using a computer, and you can change the program for different listening environments from a small, quiet room to a crowded restaurant etc.

Digital aids convert sound waves into numerical codes, similar to the binary code of a computer, before amplifying them. Because the code also includes information about a sound's pitch or loudness, the aid can be specially programmed to amplify some frequencies more than others. Digital circuitry gives an audiologist more flexibility in adjusting the aid to a user's needs and to certain listening environments and can be programmed to focus on sounds coming from a specific direction.

Hearing aids are a fairly popular solution among most age groups and users use them for about 8-9 hours a day [2].The process of getting a hearing aid is fairly simple. First, you confirm with an ENT / audiologist that you are indeed in need of one. Then a series of audiotometry tests are performed to determine the extent of damage / hearing loss incurred. Hearing sensitivity can be measured for a range of frequencies and plotted on an audiogram. Another method for quantifying hearing loss is a speech-in-noise test, which gives an indication of how well one can understand speech in a noisy environment. A person with a hearing loss will often be less able to understand speech, especially in noisy conditions. This is especially true for people who have a sensorineural loss [1] which is by far the most common type of hearing loss. A recently developed digit-triple speech-in-noise test may be a more efficient screening test.The audiologist then programs the hearing aid to amplify at an acceptable level.

## 2.2 big data in hearing aids

The working of hearing aids was covered in detail in the previous section now we shall focus on the areas where big data can be applied to help both the doctors and the patients as much as possible. We know that in order to determine the extent of hearing loss an audiotometry test will be performed. The test proceeds with a patient being made to sit in a sound proof room and being subjected to listening to a wide variety of sounds ranging from the softest possible sound they can perceive to the loudest possible. The audiologist then charts a graph to figure out the extent of hearing loss. It will look like the below graph. [figure of audiotometry here]

The problem with this process is it's still random and despite audiologists having great skill and lowering the margin as much as possible they can never be totally accurate nor can they test too much because it is physically demanding on the patient too. Big data can be a great help here. Data collected from multiple patients (with their consent) can be stored making use of technologies like apache pig , hive etc. Then when an initial audiotometry analysis has been performed we can use deep learning or simple statistical sampling to obtain a few similar cases via technologies like say apache-spark. From these cases, we can perform a more streamlined audiotometry test rather than guesswork and further accurately narrow down the loss coefficients.

After the hearing loss estimates are charted down. It is time to program the hearing aid (a hearing aid model is selected by the audiologist in accordance with the hearing loss estimates). The aid is programmed to amplify sound waves in the range where losses are observed and then various simulated environments are performed to determine the level of comfort and extent to which the hearing aid is helping and perform fine-tuning. The problem is the same as previous a very limited range of environment that may / may not be useful to the patient is observed. Using big data once again a more accurate test can be conceived. A user can be presented with the environments patients from the similar range of hearing loss faced and this can be used as a basis for fine-tuning. This process has slowly been making its way into research [4] certainly a few companies [5].

These days most people make use of digital hearing aids. As previously mentioned digital hearing aids are well equipped to make use of big data in a way they do make use of it albeit in a micro manner. The behavioural patterns of the patient are recorded like the range of volume increase or decrease in various modes, amount used etc. When the patient visits the audiologist the next time this data is analysed and then corrective changes are made towards the programming of the hearing aid. Big data can play a very big role in proactively doing so. These days most hearing aids have moved on from using a separate remote control towards making use of smartphone apps as a remote. This can be viewed as a huge enabler for big data technologies. Since mobile phones will most of the time be connected to the internet this will enable (with consent) real-time load and store of data using technologies like pig and hadoop of user environments and the current sound wave patterns and amplification used along with other useful data like if the patient is increasing or decreasing volume. Hearing aids are certainly growing smarter in the sense when a mobile communication device is brought near the aid the electromagnetic pulses

from the phone is detected by the aid and automatically switches to a phone mode, however for most of the part the user has to switch manually to other modes like theatre, noisy etc. Big data can play a huge role in automatic detection. For starters, big data can be employed to dynamically observe the fluctuations in loudness levels as well observe the fluctuations in background noise to help determine what mode should aid change to. Aside from observing fluctuations it can also compare the current scenario to those who have already encountered such scenarios under similar conditions (and with a similar hearing loss) rate and then perhaps adjust the volume/setting to a safe appropriate level. Note that this would be highly experimental and could also create more problems than it solves. As much as we have powerful machine learning algorithms like deep learning it is impossible for them to predict a solution that best suits a patient after all different patients have different problems and conditions but as it learns more and gains more data (the hallmark of deep learning to learn with more data) there will be a good chance that the algorithms will provide a solution that really suits the patient.

So far we have looked at how big data can make use of sound waves in terms of loudness background noise etc, but there is one more aspect in which big data can help us. Analysing the contents of the speech itself. It has been previously mentioned Big data in language and speech processing is a well-established topic and plenty of papers and discussions exists [3] [6]. Most speech can be well predicted these days and when we take into account that hearing aids these days can make out the direction of which the sound is coming from. Making use of this and the ability of deep learning to possibly predict parts of speech we can leverage this to accordingly increase or decrease volume. Certain words will have pronunciations that are hard to understand or have lower tones or have a high frequency to be repeated. We can take advantage of this. Also, the hearing aid can be programmed to automatically increase the volume when it detects a repetition of sentences/words either due to the patient asking the opposite person to repeat his/her sentence. This can be achieved either by listening to keywords like what, sorry, repeat yourself, etc or by analysing the sound waves and learning from that if the same pattern is being repeated.

### 2.3 Data processing and Technologies

Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data. Pattern recognition systems are in many cases trained from labelled "training" data (supervised learning), but when no labelled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning). In machine learning, pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an

input sentence, describing the syntactic structure of the sentence. From the above explanation, it becomes clear in the manner in which we can apply pattern matching for hearing aids. We could use the binary stream as a basis for calculation and from this stream try to match it to existing patterns and predict the future patterns. If any of the generated patterns are found to have speech that is hard to decipher at that range signals can be sent to the hearing aid to accordingly raise the volume of the hearing aid. Aside from that we can make use of the binary sequences and try to find the best fit pattern match, we can eliminate noise because most hearing aids these days have excellent noise cancelling technology so we can be assured that the sound stream is that of the person who is speaking to the patient. Once we get the best fit we can make a correlation between the pattern identified and the next action to be performed. The discussed methods need not only be limited to merely volume increase and decrease. They can also be applied to the fitting process as we have discussed in detail previously, after obtaining the initial estimate graph we can take the plot points [audiogram picture here] and use it to find the best-fit match of another patient and fine-tune the hearing aid accordingly saving a lot of time and effort and allowing the entire experience to be pleasant and more productive.

Apache hadoop, hive etc are just data warehousing software, used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. We have previously seen how data is being made available for us the next logical question that comes to mind is how do we store it and using what. The answer to what is somewhat easier to explain. As mentioned previously we can make use of sound waves which gets converted to numerical codes. Mobile phones used to control the hearing aids which have access to the internet can make use of hadoop framework. We can send this data to some API which then uses map reduce to accordingly save it to a location which represents that pattern. While looking for a pattern the analysis made at real time can be used to query the API which will use map reduce to obtain all patterns relevant to the range of hearing loss and then use pattern matching to figure out best fits and suggest feasible solutions. Audiologists can make use of this in the same manner only instead of making use of the API via mobile phones they can do it conventionally by a computer which will allow for stronger processing.

### 2.4 Section Summary

Hearing aids are the most assistive technology a person with hearing impairments could receive which do not require surgery. They are fast becoming an important industry especially with the rising problem of hearing loss. The process of obtaining a hearing aid is simple and straightforward and for a long time it remained static. However with the advent of big data the world of hearing aids has been shaken from its static foundations and is undergoing a paradigm shift in the same manner mobile phones evolved to the current smart model. Right from the process of fitting to using the hearing big data can be applied to almost every stage. The entire

process is still in its infancy and there is a huge scope for further development. Pattern matching , deep learning all the concepts are only the tip of the iceberg there are many algorithms and designs that can Be implemented. Thus there is a huge market both towards improving the current crop of hearing aids available and as well as creating jobs.

### 3 BIG DATA IN CLOSED CAPTIONING

#### 3.1 Introduction

The importance of video in today's world cannot be stated enough. in the field of education, more and more classes are being shifted to the online model, professors are moving towards keeping their lecturers in places like udemy, youtube, pluralsight etc because there they are allowed the total freedom to take the course in their own direction without the constraints of time and classroom size. Certain sites like youtube even offer a source of remuneration. Thanks to the advancements in technology like smartphones and cheaper internet the general public is not satisfied with just listening to the audio, they want to watch the video and embrace the whole experience and artists have contributed to it by making their videos so colourful. TV is ever present with people preferring to watch news comfortably and be updated on the go rather than rifle through a newspaper. Even in communication, there is a paradigm shift from normal audio based conversations heading towards video calls, facetime, Whatsapp video chat are just a few of the most popular options available.

Now that an accord has been reached on how popular the video format is becoming, its time to talk about how difficult it's becoming for people who have hearing disabilities to cope with this changing world. These are people who are struggling to understand communication face to face and now are tasked with trying to understand something available on a digital platform and something they may not even have the power to ask for a repeat in case they misheard. the problem with digital media is mainly the distortion that comes with it. If it is too loud it becomes illegible to understand if it's too soft it is not loud enough to understand and the middle ground isn't much of a help more often than not. Aside from videos hearing impaired patients also find it difficult to perform most of their daily life activities like attending a class etc.

So what solution do we have that is capable of solving most of the above problems. It can be observed that the community has gone about finding different ways to solve their problems. Sign language translators, lip reading etc. They are all convinient methods of getting by but the problem with them is as good as they are they are way too situational. A more reliable method would be that of closed captions. Note that the art of captions is one that already exists and is made mandatory by governments to make such resources available on request and in general people are very receptive towards these technologies and often go out of their way to enable them. For students perhaps the most common use of captions is the CART(Communication Access Real Time) systems, where an individual types a captioning of what the teacher is saying and the students view this in real time there will always be a delay of a few microseconds but it is still extremely useful. In the case of tv, most channels already have implemented a subtitle system. Pre-recorded

shows already have subtitles generated in advance and they are displayed. In case of live settings, the same text from the teleprompter is used as a subtitle or a cart like a system is used where a person types in the text being spoken and it is displayed a few seconds later.In the case of online videos, most content providers have provided the option of loading the subtitle file. The problem now lies in generating/obtaining the subtitle files. Good Samaritans always exist and for important videos more often than not the uploader or someone else will generate their own subtitles. The society is slowly recognising the need for subtitles and in general, you will find the subtitles of the files you are looking for if you look hard enough. There also exist professional agencies who will create subtitles either for free or for a price, though in general, the free ones are already hard at work.

The issue at hand today is generating good quality subtitles and fast. The traditional method of generating subtitles by having a person listen and then type the subtitles is fine but it is too slow and cumbersome. It would be faster to have a computer generate the subtitles and have a human change/modify errors. Youtube has its own automatic subtitle generation, Facebook too is developing a similar feature on the same lines. Watson IBM have developed an API that leverages the power of Watsons AI to generate subtitles.

*3.1.1 Manual caption generation.* Manual caption generation as previously stated is a very tedious job it involves a person listening to the audio and then accordingly typing out the captions and storing them for later use. Big data can play a significant role in making thing easier for both the generator and the consumer. First lets deal with the generator. In terms of manual caption generation, there is not much that can be done to aid the person unless he switches to an automatic caption generation system. A few simple steps, however, can make his/her lie much simpler. Using big data the video could be analysed for segments which have very common phrases / repeated segments and then these segments could be stored using map reduce. the next could be using map reduce find if there were translations available previously and generate the captions or store it till the user enters the caption. There are multiple benefits of such a method, perhaps the most important benefit would be it serves as an error detection mechanism improving the subtitle accuracy. Another benefit is it serves as a great training tool for supervised and semi-supervised learning algorithms. For the consumers, the biggest problem is finding subtitles. Using Apache Hadoop etc we can create an API that acts as a centralized database. However, this is an initiative that should be encouraged by the government. This is a problem that affects all citizens and having the government take care of it adds responsibility and accountability.

*3.1.2 Automatic caption generation.* Automatic caption generation has made huge strides in the recent years thanks to the development in the field of AI, semantic web, statistical models etc. In the case of automatic caption generation, the process is amazingly straightforward. First, you select the video for which you want to generate. then you upload/send the data to the respective API. The API then generates the captions you need which you can then embed into the video or load into the video.The favourite method seems to be that of youtube API which gives a very high accuracy rating of around 94% [cite TheDeafCaptioner2017]. Aside from youtube, there exists IBM Watson which boats of one of the most

powerful deep learning features. It should be noted that the youtube method is not exactly real-time sure it can work in real time with a delay of a few seconds but it requires the video to be clipped and sent at regular intervals in order for it to be real time to return a caption response with a slight delay. That being said it doesn't mean that youtube cannot do it in real time its a feature that has not been made available to the public yet, to clarify the speech to text conversion happens at a real-time rate it is just the communication between sender and server that will take some time. Youtube API much like Watson makes use of powerful and complex deep learning neural networks. While the exact process has not been made to the public quite understandably so the general concept will be explained along with an additional Hidden Markov model method.

**3.1.3 Automatic caption generation markov model.** First, we shall explain a simple and relatively straightforward hidden Markov model a simple but powerful AI algorithm. A formal definition from Wikipedia that helps in understanding the concept of Markov models is given as " In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters, while in the hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. The adjective hidden refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a hidden Markov model even if these parameters are known exactly." [markov model figure] The thought process that follows the following explanation is that of a semi-supervised learning. The training data will compose of just words(audio) with their respective speech tag. Also, the training data will comprise a good set of perfectly captioned videos. More than the videos it is the caption files that are important. The individual words file merely provide us with a corpus file of what words to expect. it is the caption files however that allow us to build the Markov model around which the entire caption generation process will be built upon. Now that we have our corpus file we can begin analysis on the caption file to build the model. What we will be doing is using a very simple probabilistic equation, The probability of observing a sequence  $Y = y(0), y(1), \dots, y(L-1)$  of length  $L$  is given by,

$$P(Y) = \sum_X \frac{P(Y/X)}{P(X)}$$

To explain this in a more understandable manner we are looking for relationships between words. For example, when a person is introducing themselves a common sentence is "Hello my name is xyz" or "hello I am xyz". Another assumption we will have to make at this point at least with respect to the current Markov method is that we will work with more complete words that span at least 3 or more syllables so that the model can detect them better. With that, the previous sentences become "hello name xyz" and "hello xyz", from this reduction we can now get a new relationship, two of them in fact. the word hello leads to the follow up of the word name and this leads to the follow up of xyz whatever and hello

leads to xyz. Now this relationship will be remembered. Now we can also make use of this concept but towards sentences. First, we will have to determine via some statistical method what would be the best average length [ time duration] for sentences. Now we can use the previous principle to observe if there exist any relationship between sentences and remember them. Now we get on to processing the data to generate the captions. First, we have to eliminate noise from the audio. Then we need to split it into  $n$  appropriate sections [ $n = \text{total length of video}/\text{time of average sentence calculated previously}$ ]. Now using MapReduce we can quickly determine how many segments already have a translation [ie find if there are any perfect/reasonable matches] and obtain the captions for those. If there are none available then we go to the next stage of getting captions word by word. After converting it to a data format appropriately [this has been discussed this previously in the hearing aid section, another alternative manner will be discussed shortly]. Now we determine the first word, again using map reduce we can match the word to a list of probable similar sounding words. To obtain the actual word the magic of probability comes into play. remember the Markov model we had discussed earlier, from it we can calculate the probabilities of a word occurring as the first word and accordingly make a very educated guess on which word is the right word. Following that using the Markov model once we can determine what the next word could be after first narrowing down the suspects we can use the model to infer what the next word could be. that is, given the current word what is the probability that the next word will be the selected word. This way we go about generating captions for every word in the sentence. After this is done for all sentences we can generate a caption file for the entire duration of the video. This method is likely to have a lot of errors since the Markov model described here is rather simplistic therefore the onus will have to be on the administrator to go over the captions generated and accordingly make the corrections. This will not only be useful to the people with disabilities but also help improve the accuracy of the system on the whole. Studies have been carried out and more complex implementation of the same idea has been performed with a rather high accuracy rate can be found here [ cite Markov model paper].

**3.1.4 Automatic caption generation Deep neural network model.** Now the concept of a neural network will be explained. A neural network is based on the same idea of how our brains work. A collection of neurons for information processing and to model the world around us. A very brief explanation would be a neuron sums all the inputs and if the resulting value is higher than a specified threshold it fires.[neural net picture]. The above configuration is called a perceptron. It has  $n$  inputs and  $n$  weights are real numbers and can be positive or negative. The perceptron consists of weights, summation processor and an activation function. the inputs are multiplied by the individual weights and the summation of all of these is passed to an activation function, we will make use of a step activation function which fires 1 if above the threshold a 0 otherwise. We need to train the perceptron now. This essentially means modifying weights after observing the inputs such that the activation function fires correctly. For all inputs,  $i$ ,  $W(i) = W(i) + a^*(T-A)^*P(i)$ , where  $a$  is the learning rate, here,  $W$

is the weight vector. P is the input vector. T is the correct output that the perceptron should have known and A is the output given by the perceptron. When an entire pass through all of the input training vectors is completed without an error, the perceptron has learnt. A deep neural network is thus a collection of perceptrons or to be more accurate it is a multiple layered architectures which compromises of an input and output layer and in between them multiple hidden layers. [multilayer network image] From the image we can observe Each input from the input layer is fed up to each node in the hidden layer, and from there to each node on the output layer. We should note that there can be any number of nodes per layer and there are usually multiple hidden layers to pass through before ultimately reaching the output layer. But to prepare this we require a learning calculation which ought to be able to tune not as it were the weights between the output layer and the hidden layer but moreover the weights between the hidden layer and the input layer. Clearly, it becomes obvious that we will need to tune the inputs between the input layer and hidden layer for this we shall make use of a technique called backpropagation which essentially means we carry the error to the next stage of input and then use these errors to modify the input stage of every layer. to be brief we can summarize it as follows We present a training sample to the neural network (initialised with random weights). Compute the output received by calculating activations of each layer and thus calculate the error. Having calculated the error, we readjust the weights (according to the above-mentioned equations) such that the error decreases. We continue the process for all training samples several times until the weights are not changing too much [cite pokarna].

Now that we have a good understanding of how neural networks work we shall look into how ASR happens via neural networks. we shall go about this step by step. The first problem will be converting sound to bits. We have previously seen digital hearing aids automatically convert sound waves into numerical codes and have extended this concept in multiple places. We shall now go into this a little more in detail and provide an insight into how this could happen. Consider the below wave of sound [sound wave] This can be represented on a graph as a simple expression of height vs time. The annotations towards height can be amplitude, frequency whatever a person chooses that suits their purpose best. However, note that this is a bit too scattered and not very uniform which is understandable because over digital media the voice can break. Let us attempt to smooth this signal using one of the many transformation algorithms available [ Nyquist theorem for example]. The result will look like this [figure smooth hello]. Recall it was mentioned that we can simply take the graph as a function of height vs time, so to help visualize this consider the [figure sampled graph]. To represent this in a text format it would look like [ 100,-20,30,89,789,-400,.....345] where each value is a measure of height over a designated unit of time. We need to be careful about how we select this unit of time. The idea is that different syllables have a different pitch we want to exploit that. Hence we want to select a unit of time such that each unit possibly corresponds to one letter. The [figure] gives an insight into how the overall model should look. So after we have trained the network to recognise each letter based on value/set of values we proceed with feeding the current input stream and thereby obtaining each letter. We can then pair these letters to form

a word. Note that there is a high chance that many letters could be repeated, decisions must be made based on observations if we need to replace repetitions. Once we have obtained these words there is a possibility of spelling errors using an auto-correct program [there are many good algorithms available on the internet] we can then use the Markov model we previously explored to verify its correctness [ if the word generated fits the model observed]. To save some time we could use map-reduce again to find out if similar patterns exist before and then use it accordingly to determine if there were similar instances before.

*3.1.5 Using Image Processing to generate captions.* It is well known how powerful OCR is

## 4 CONCLUSION

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

## REFERENCES

- [1] AG Bell Academy. 2017. causes of hearing loss. *ag bell* 1, 1 (Dec. 2017), 1. <http://www.agbell.org/learn/hearing-loss-explained/causes-of-hearing-loss.aspx>
- [2] J Am Acad Audiol. 2017. Hearing Aid Use and Mild Hearing Impairment: Learnings from Big Data. *jaaa* 1, 1 (Sept. 2017), 1. <https://www.ncbi.nlm.nih.gov/pubmed/28906244>
- [3] I-Hsin Chung. 2017. Parallel Deep Neural Network Training for Big Data on Blue Gene/Q. *jaaaa* 1, 1 (Dec. 2017), 1. <http://ieeexplore.ieee.org/document/7013048/?reload=true>
- [4] peter nordquist. 2017. Quality assurance in health care based on Big Data. *jaaa* 1, 1 (Dec. 2017), 1.
- [5] phonak. 2017. phonak uses big data. *phonak* 1, 1 (Jan. 2017), 1. [https://www.phonakpro.com/content/dam/phonakpro/gc\\_hq/en/resources/evidence/field\\_studies/documents/fsn\\_autosense\\_os\\_big\\_data.pdf](https://www.phonakpro.com/content/dam/phonakpro/gc_hq/en/resources/evidence/field_studies/documents/fsn_autosense_os_big_data.pdf)
- [6] Bjrn W. Schuller. 2015. Speech Analysis in the Big Data Era. *jaaaaa* 1, 1 (Dec. 2015), 1. [https://link.springer.com/chapter/10.1007/978-3-319-24033-6\\_1](https://link.springer.com/chapter/10.1007/978-3-319-24033-6_1)
- [7] Wikipedia. 2017. Hearing loss. *wikipedia* 1, 1 (Dec. 2017), 1. [https://en.wikipedia.org/wiki/Hearing\\_loss#Causes](https://en.wikipedia.org/wiki/Hearing_loss#Causes)

## 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

[2017-12-10 13.49.31] pdflatex report.tex

=====

## Compliance Report

=====

```
name: Rahul Velayutham
hid: 232
paper1: 2017-10-29 100%
paper2: 100%
project: in progress 40%
```

```
yamlcheck
-----
```

```
wordcount
-----
```

```
6
wc 232 project 6 6517 report.tex
wc 232 project 6 6373 report.pdf
wc 232 project 6 666 report.bib
```

```
find "
-----
```

92: Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data. Pattern recognition systems are in many cases trained from labelled "training" data (supervised learning), but when no labelled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning). In machine learning, pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

126: First, we shall explain a simple and relatively straightforward

hidden Markov model a simple but powerful AI algorithm. A formal definition from Wikipedia that helps in understanding the concept of Markov models is given as " In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters, while in the hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. The adjective hidden refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a hidden Markov model even if these parameters are known exactly." [markov model figure]

- 128: The thought process that follows the following explanation is that of a semi-supervised learning. The training data will compose of just words(audio) with their respective speech tag. Also, the training data will comprise a good set of perfectly captioned videos. More than the videos it is the caption files that are important. The individual words file merely provide us with a corpus file of what words to expect. it is the caption files however that allow us to build the Markov model around which the entire caption generation process will be built upon. Now that we have our corpus file we can begin analysis on the caption file to build the model. What we will be doing is using a very simple probabilistic equation, The probability of observing a sequence  $Y = y(0), y(1), \dots, y(L-1)$  of length of  $L$  is given by,  $P(Y) = \prod_{i=1}^L P(y_i | y_{i-1})$ . To explain this in a more understandable manner we are looking for relationships between words. For example, when a person is introducing themselves a common sentence is "Hello my name is xyz" or "hello I am xyz". Another assumption we will have to make at this point at least with respect to the current Markov method is that we will work with more complete words that span at least 3 or more syllables so that the model can detect them better. With that, the previous sentences become "hello name xyz" and "hello xyz", from this reduction we can now get a new relationship, two of them in fact. the word hello leads to the follow up of the word name and this leads to the follow up of xyz whatever and hello leads to xyz. Now this relationship will be remembered. Now we can also make use of this concept but towards sentences. First, we will have to determine via some statistical method what would be the best average length [ time duration] for sentences. Now we can use the previous principle to observe if there exist any relationship between sentences and remember them. Now we get

on to processing the data to generate the captions. First, we have to eliminate noise from the audio. Then we need to split it into n appropriate sections [ n = total length of video/time of average sentence calculated previously]. Now using MapReduce we can quickly determine how many segments already have a translation [ ie find if there are any perfect/reasonable matches] and obtain the captions for those. If there are none available then we go to the next stage of getting captions word by word. After converting it to a data format appropriately [ this has been discussed this previously in the hearing aid section, another alternative manner will be discussed shortly]. Now we determine the first word, again using map reduce we can match the word to a list of probable similar sounding words. To obtain the actual word the magic of probability comes into play. remember the Markov model we had discussed earlier, from it we can calculate the probabilities of a word occurring as the first word and accordingly make a very educated guess on which word is the right word. Following that using the Markov model once we can determine what the next word could be after first narrowing down the suspects we can use the model to infer what the next word could be. that is, given the current word what is the probability that the next word will be the selected word. This way we go about generating captions for every word in the sentence. After this is done for all sentences we can generate a caption file for the entire duration of the video. This method is likely to have a lot of errors since the Markov model described here is rather simplistic therefore the onus will have to be on the administrator to go over the captions generated and accordingly make the corrections. This will not only be useful to the people with disabilities but also help improve the accuracy of the system on the whole. Studies have been carried out and more complex implementation of the same idea has been performed with a rather high accuracy rate can be found here [ cite Markov model paper].

passed: False

find footnote

---

passed: True

find input{format/i523}

---

4: \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: figure and below may be used improperly
```

```
79: The working of hearing aids was covered in detail in the previous  
section now we shall focus on the areas where big data can be  
applied to help both the doctors and the patients as much as
```

possible. We know that in order to determine the extent of hearing loss an audiometry test will be performed. The test proceeds with a patient being made to sit in a sound proof room and being subjected to listening to a wide variety of sounds ranging from the softest possible sound they can perceive to the loudest possible. The audiologist then charts a graph to figure out the extent of hearing loss. It will look like the below graph.

WARNING: code and below may be used improperly

135: Now that we have a good understanding of how neural networks work we shall look into how ASR happens via neural networks. we shall go about this step by step. The first problem will be converting sound to bits. We have previously seen digital hearing aids automatically convert sound waves into numerical codes and have extended this concept in multiple places. We shall now go into this a little more in detail and provide an insight into how this could happen. Consider the below wave of sound [sound wave]

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 233  
non ascii found 232  
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 8211
```

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

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

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

55: \newline

57: \newline

81: \newline

83: \newline

85: \newline

87: \newline

93: \newline

95: \newline

97: \newline

111: \newline

113: \newline

116: \newline

124: \newline

```
127: \newline  
129: \newline  
134: \newline  
passed: False  
cites should have a space before \cite{} but not before the {
```

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

```
passed: True
```

# Analyzing everyday challenges of people with visual impairments

Tousif Ahmed  
Indiana University  
150 S Woodlawn Avenue  
Bloomington, Indiana 47405  
touahmed@indiana.edu

## ABSTRACT

People with visual impairments face varieties of problem in their daily lives. Nowadays, modern technology especially camera-based technologies are helping people with visual impairment in their everyday tasks ranging from daily household activity to navigation. Users are using camera based applications where they are sharing photos and asking questions. Based on the asked question and shared photo, automated tools or human crowd workers are helping the visually impaired people in their tasks. By exploring the questions, it is possible to understand the problems and challenges of people with visual impairments. However, the volume of such data makes it impossible to analyze the questions manually. Big data analytics could help us to understand the challenges of people with visual impairments. To understand the challenges, we analyzed the VizWiz data set which contains more than 33,500 questions asked by people with visual impairments. In this paper, we report on the data and shed light on the challenges.

## KEYWORDS

E534, HID 237, Big Data, Accessibility Issues, People with Visual Impairments

## 1 INTRODUCTION

People with visual impairments face a variety of problems in their daily lives and need assistance. They need assistance with detecting objects, identifying money, navigation, transportation, household activities, cooking, and various other activities. Sighted person on rely on vision on so many things that it is almost impossible to visualize and understand the problems of people with visual impairments. Although there are variety of tools available to simulate the challenges and experiences of people with visual impairments, the challenges of people with visual impairments is not well understood. To help visually impaired people with technologies, we need to understand their problem first.

One possible to understand the challenges of people with visual impairments is qualitative analysis or ethnographic studies with people with visual impairments. Simply, researchers can follow or conduct interviews with people with visual impairments. Although qualitative studies are widely accepted research methods, it has limitations. Specially to understand the problems of people with visual impairments, qualitative studies have severe limitations. As the challenges vary with the experiences of people with visual impairments, these studies can not capture or depict the whole picture. Besides, these studies are very expensive and need ample

human effort. Therefore, we need a better way to understand the challenges of people with visual impairments.

Big data analytics could be a potential alternative. To understand how big data can help people with visual impairments, we need to understand the background first. Nowadays, people with visual impairments uses different technologies for their problems. A wide range of technologies such as talking watch, braille reader, navigation helper are available in the market to help the visually impaired in their daily tasks. Since the introduction of smartphone, smartphone based applications gained huge popularities among people with visual impairments. Now, mobile and smartphone applications like Seeing AI [7], AiPoly [2], LookTel [9], and other such camera based applications are helping people with visual impairments in object recognition, face recognition, color detection, human emotion detection, activity recognition, and other such tasks that was not possible before. Figure 1 depicts an example from Seeing AI which shows that how camera based applications are helping people with visual impairments by describing nearby person's activity (Figure 1a) and their facial emotions (Figure 1a).

[Figure 1 about here.]

Most of the camera assisted assistive applications works in one simple way. The user uploads an intended photo and asks a question about that. Applications have its simple iq engine which tries to answer the problem first. If it's not able to answer that question, it shares the questions and images with the user's friends and family. Sometimes, the image is shared with a web based human worker. This crowd worker is essential for such system, because the iq engine is not sophisticated yet. We can not completely trust the automated approaches. Besides, visually impaired user's can not efficiently take photos. Sometimes they point at totally wrong objects or items, sometimes they share blurry photos, and even sometimes the question does not match with the photos [3, 5, 6]. Therefore, to give a correct answer of the questions, technologies require human intelligence. Questions and answer based applications like TapTapSee [8] and VizWiz [3] uses this approach. LookTell [9] and BeMyEyes does not have any automated approach, it directly broadcasts the video feed to the volunteers and volunteers answer their questions. Some applications are trying to move towards the fully automated approach, however, due to the limitations of automated approaches they did not gain much popularity yet.

Human based systems have privacy issues, because these users are uploading their photos which may contain sensitive information. Often they ask about medical information, their address, and various other sensitive information which can be exploited by the malicious crowd workers. Even sometimes, the users shares their credit card image and asks the system to about their credit card

information which have severe privacy and security implications. Moreover, cameras and images shared by them can be extremely risky for people with visual impairments, because often they do not know the contents of the photo. Photos can be uploaded in error, sensitive data can be shared unintentionally. Ahmed et al. [1] reported a scary story of one of the VizWiz users, who accidentally shared her naked photo with a crowd worker. Such evidents suggests that such systems has severe privacy and security implications. However, visually impaired needs such tool in their daily lives. Therefore, the ideal solution would be an completely automated approach. However, to design a flawless system we need to improve the existing tools first and we need to understand the challenges first.

The challenges of people with visual impairments can be easily understood from the images uploaded and questions asked by the user. Although it is extremely difficult for people with visual impairments to take a good photo, still they are using these tools because of their challenges. Therefore, the data uploaded in these applications are probably a good way to understand the challenges. However, due to the volume of the data it is not possible to manually identify the problems. Therefore, big data analytics can be helpful in this context to understand the challenges of people with visual impairments. However, due to privacy issues all but one data sets are not publicly available.

In this paper, we analyzed the VizWiz data set containing more than 35000 images and questions [3]. Based on the questions asked, we tried to categorize their problems which eventually help the researchers to design and develop a fully automated system. Previous researches [4] explored the same problem with the same data set. However, they only explored 1000 images and performed a qualitative study and identified four categories of problem. Since manual analysis is not possible on 35000 data, we used big data tools to automatically analyze the questions and images. In this paper, we report on analysis performed and the visual challenges of people with visual impairments.

## 2 METHODOLOGY

In this section, we discuss the methodology for identifying the challenges of people with visual impairments.

### 2.1 Data Set

We used VizWiz data set which is the only available data set in this category. VizWiz is an iPhone application that allows visually impaired users to get quick responses of their challenges. The app tries to find an answer by using automatic IQ engine and anonymous crowd workers [11]. VizWiz was released in May, 2011.

VizWiz application helped more than ten thousand users by answering more than 100,000 questions. However, they only shared half of their data from those participants who gave consent. Therefore, around 50,000 data are available for the researchers. However, the researchers removed around 6,000 data due to sensitive contents in the images. The rest of the 43,543 images were made public. All images and questions redundantly checked and anonymized. We downloaded this data set for research purposes in May. Recently, the data is not available to download. Therefore, we urge the instructors to not distribute the data.

Based on the images shared and questions uploaded, we found only 33,580 images and their related questions. The questions were shared in json format and images are shared in a compressed directory. The questions data set have three columns which I described next:

- **image:** The name of the image file.
- **private:** If the image is marked as private.
- **question:** Asked questions of the user. Some questions are missing.
- **response:** Each question can have multiple responses. As mentioned earlier, some questions were tried to answer using the IQ engine and some questions were sent to the web workers. For each question, there can be one to 11 responses. However, on average three responses were received. The distribution of the responses shown in Figure 2. From the figure, we can see that most of the questions either received three or four responses.

[Figure 2 about here.]

### 2.2 Data Cleaning and Preparation

Since this data was collected from a research group, the data is very clean. We did not need any further cleaning except we discarded the private column. Since the researchers did not share the private images, therefore, all the rows in private columns shows false. Therefore, this column does not add any values to our analysis. We also noticed that lot of questions are missing, but an image is available. We can safely assume that these images were asked without questions and the users assumed that the images are self describing. Since the images can be interesting, therefore, we still kept the questions and labeled those questions as 'NoQues'. We used 'pandas' for storing the data. We also uploaded the images in the specified Google Drive folder.

### 2.3 Data Analysis

We performed analysis on both the questions and image data sets. The image data set only used to detect the blurred images. However, we rigorously analyzed the questions to identify various issues of people with visual impairments. In this section, we mainly discussed the text analysis methods. The full analysis can be found in 'question\_analysis' jupyter notebook. The image analysis can be found in 'image\_analysis' notebook.

**2.3.1 Question Analysis.** To understand the challenges of people with visual impairments we performed unigram, bigram, and trigram analysis. Based on the analysis, we identified several issues which is presented in the results. The process of identifying the challenges is discussed in next section.

## 3 RESULTS

In this section, we present our results that we identified from the analysis:

### 3.1 Identifying the challenges of people with visual impairments

The questions asked by people with visual impairments explains some of their challenges in their daily lives. Whenever they are

facing issues, they are asking questions in VizWiz. Therefore, the questions asked could give us some insights about their challenges.

[Figure 3 about here.]

To understand the challenges, we first calculated the frequency of the words. There are around 4500 unique words in the questions. The most frequent 50 words is shown in Figure 3. If we closely examine the words, we can see that the most frequently used word is ‘what’. ‘What’ appeared 22793 times which is approximately 70% of all of the worlds. The second and third most frequent words are ‘this’ and ‘the’. Since, this is a set of questions, therefore, all the above words are justifiable. Although, ‘what’ is somewhat giving us an indication that users are asking about objects or subjective questions mostly, ‘this’ and ‘the’ is not adding that much value. Next, we performed the same analysis by removing the most commonly used words in English. That unigrams gave us some additional insights. The list of most frequently used interesting words can be found in Figure 4. If we remove the commonly used words, then for the majority of the questions had no questions. Those questions were asked by just uploading the photos. We assume that the users thought that the app could automatically answer those questions. Other three most frequently used words are ‘color’, ‘tell’, and ‘please’. Among these three the most interesting is ‘color’. Combination of ‘what’ and ‘color’ indicates that people with visual impairments faces issues with color detection, and often they ask the workers about the color of the objects and items. Therefore, we found **color detection** problem of people with visual impairments from the analysis. If we just consider the nouns and pronouns from the 30 most frequently used words, we find ‘box’, ‘picture’, ‘color’, ‘screen’, ‘shirt’, ‘bottle’, ‘flavor’, ‘brand’, ‘coffee’, ‘label’, and ‘product’. From this keywords, we can safely assume three other problems: they face issues with screens (screen), there are issues with objects (brand), and the users face issues with reading labels. Therefore, from the initial analysis we found four problems that people with visual impairments regularly face: **color detection, object detection, reading screens (mobile/ computer), and reading labels.**

[Figure 4 about here.]

After checking the most frequently used words, we explored the most interesting pairs of words. If we check the bigrams (Figure 14 and 15 in Appendix), it gives confidence of our identified problems. The most frequently used two words are ‘what’ and ‘this’ which suggests that most questions were asked to identify the object. Therefore, people with visual impairments definitely face problems with detecting objects. ‘What’ and ‘color’ also suggests that users face color detection problem frequently. If we check the bigrams of most frequently used interesting words (Figure 15), we find some additional insights. If we ignore ‘NoQues’, then we again see color detection and computer screen reading problem. However, now we can find another interesting pairs of words ‘look’ and ‘like’. This pair indicates a subjective question, where the user is asking how the user is looking like. This identifies another challenges of people with visual impairments **Impression Management**. Another interesting common pair of words are ‘long’ and ‘cook’ which indicates reading label issues, however this can be a household activity issue. The trigrams also gave us some new interesting insights (Figure 16). Most of the trigrams confirms above mentioned challenges,

however, there are some new issues. One interesting trigram is ‘display’, ‘treadmill’, and ‘tell’ which indicates the health fitness related issues or **Health Management Issue**. Due to the accessibility issues in health monitoring and fitness monitoring issues, they can not manage health effectively. Therefore, the users often seek help for reading the display. Another interesting three words are ‘pregnancy’, ‘test’, ‘show’ which can also be put into health Management category. However, this seems a private information, but still people with visual impairments have to share this information due to their visual challenges.

### 3.2 Challenges

Based on the rigorous analysis, we identified some challenges of people with visual impairments. In this section, we discuss the challenges:

**3.2.1 Object Detection:** The most frequently asked question in VizWiz is ‘What is this’ or ‘What is that’. ‘What’ appeared more than 22,000 questions. Among those 22,000 questions around 7,000 questions are ‘What is this?’ and ‘What is that?’. People ask variety of object detection questions ranging from everyday objects to personal objects. Some examples of object detection problem is shown in Figure 5. By manually analyzing some photos, it seems most of them are related to household activities. Therefore, with better tools it is possible to detect the objects.

[Figure 5 about here.]

**3.2.2 Color Detection:** Another most frequent problem that people with visual impairments face is to detect colors. Most of the time they use VizWiz to identify colors of their cloths, items, foods, and others. Some examples of color detection is shown in Figure 6. Based on the images, automatically detecting the colors seems a challenging task. Because, if we examine figure 6 we can see in the image there can be other objects. Automatically detecting the object of interest will be difficult. For example, in the right most photo the user is asking about the color of the dress in hand, however, there are other objects visible in the photo. Therefore, identifying the color automatically will be challenging.

[Figure 6 about here.]

**3.2.3 Reading Screens:** Nowadays, people with visual impairments use smartphones and computers. They use screen reading software which generates synthesized speech to relay the information from screen. However, sometimes these software fail and visually impaired need to seek help from crowd workers. Another issue is the accessibility issues of CAPTCHA, people with visual impairments struggle with CAPTCHA. Therefore, they seek people who can read the CAPTCHA for them. Some examples of reading screen problems are shown in Figure 7.

[Figure 7 about here.]

**3.2.4 Reading documents or labels:** Another obvious challenges of people with visual impairments is reading documents. The paper documents are not often accessible and people need help from others to read that. People might use scanners to read documents, however, scanning documents can be time consuming. Especially, for scanning food or medicine labels can be difficult. Therefore, participants seek help to read labels for them. Figure 8 shows some

examples of reading issues. However, there can be potential score for technology for this types of problem. If the user is asking for reading helps, a simple OCR can help. However, OCR might not work well with food labels. One suggestion could be for food related reading question, the system could look for barcode and identify necessary information.

[Figure 8 about here.]

**3.2.5 Impression Management:** Based on the analysis, we explored that managing impressions can be challenging. As a social norm, we often present our better selves to others by wearing consistent dresses. For example, we do not want to present ourselves in social places in such way that may misrepresent ourselves. Some words that we found in the questions are ‘look’, ‘like’ which we assume that users are asking to understand their appearance. Therefore, impression management for people could be challenging. Sometimes, the questions can be appearance related. Some examples of impression management challenges is shown in Figure 9.

[Figure 9 about here.]

**3.2.6 Health Management:** Health management is important for everyone. However, people with visual impairments face lot of challenges to maintain healthy behavior. They struggles to cook, therefore, they need to eat outside or eat packaged foods. They can not read the package’s well, so miss the nutrition info. Managing medicine can be issue. Some other issues can be attributed to visual representation of results. For examples, weight scales show visual weights, pregnancy scales convey visual feedback, health monitoring instruments like treadmill convey visual information. All these visual information makes it difficult for managing health issues. Therefore, health management can be challenging. For that reasons, people with visual impairments often ask such applications to help them with various visual indicators in health and fitness. Figure 10 shows three different health realted issues of people with visual impairments. Figure 10a depicts the issues of medicine management, users often can not identify the required medicine. Figure 10b shows asking the result of pregnancy test, which can be sensitive. Figure 10c asking questions about the weight of the user. Since, such applications can forward these questions to friends and family members all these images can be sensitive. However, technology can potentially address this issue by automating the responses.

[Figure 10 about here.]

**3.2.7 Taking Photos:** Like sighted people, visually impaired people also wants to take photos. However, taking photos are challenging since the users can not seen the image. Therefore, they often struggle to take photos. The irony of applications like VizWiz is that these services require a challenging task to solve other challenges. Although none of the questions mention anything about taking photos, the responses of the web workers illustrates the photo taking challenges of people with visual impairments. Around 4000 images have been detected as blurry and not understandable by human workers. Apart from blurry images, sometimes photos can be out of focus and misplaced.

[Figure 11 about here.]

Figure 11 depicts the some not understandable photos taken by people with visual impairments. However, such images takes resources and often cost money. If the system can early detect such images and prevent those images from sending then it can save resources. Misplaced or blurry photos can be early detected. Another potential scope of technology is to automatically fix the blurry images.

We identified various challenges of people with visual impairments. There can be other challenges, however, from the VizWiz data set these seven seems some major problems. We also discovered that there can be privacy issues with the shared images (i.e., pregnancy test results) and such data need to be handled carefully. Although existing services require manual efforts, technology has various scopes to help people with visual impairments. Due to poor quality of images, such system may consume significant user resources and early detection of the quality of images can save the resources. In the next section, we discuss one such approach and the evaluation of the approach using VizWiz data set.

### 3.3 Automatically Detecting Blurry Images

We have already give some examples of the struggles of taking photos by the user. Often their photos are out of focus and blurry. Using OpenCV, we can detect blurry images. From the web workers responses we have an estimation that some photos are very blurry and can not be recognizable by human. If the system can early detect the blurry photos and asks the user to retake the photos it could reduce human effort. In this analysis, our task is if we can automatically identify the blurred images. The ‘Image Analysis’ jupyter notebook shows some the analysis that we performed in this section.

**3.3.1 Estimation of Ground Truth Data.** We set up the ground truth from the web workers responses. If any of the web workers mentioned that the image is blurry, then we set the image as blurry. From that, we found a list of 3580 images which can be considered as blurry. We then divided the data frame into two different sets: blurred set and not blurred set.

**3.3.2 Detecting Blurry Photos.** We followed pyimagesearch’s tutorial to detect the blur images [10]. Following that tutorial, we used variance of the Laplacian to detect the blurred images. Then, we run the algorithm on 33,580 images.

**3.3.3 Calculating F1 score.** We made an assumption for the accuracy of blur detection. If we consider the real case scenario, if the user need to take a photo more than once to avoid blurring that is not a problem. Although, they have difficulties of taking photos but it is still possible to take a better photo and there is no cost of taking photos. However, if we send a blurry photo to web worker it wastes resources. The system need to pay the web workers for their tasks and the system somehow charges that money to the users. Therefore, taking a blurry photo is costly. Therefore, for such a system it is better to be some false positives than false negatives. Therefore, this system tries to reduce the false negatives. Hence, we tried to improve the recall. However, too much false positive can affect the usability of a system. Therefore, Our target is to find the best accuracy over blurred images minimizing false positive rates. F1 score will help us to find a correct threshold. Our initial threshold of 150 gave us F1 score of 21.36%.

**3.3.4 Identifying a good threshold.** We run the algorithm with various thresholds. The F1 score graph against various thresholds did not improve the accuracy. Figure 12 shows the accuracy of blur detection.

[Figure 12 about here.]

**3.3.5 Implications of result.** The poor accuracy of the blur detection algorithms depicts some problems of real world data set. Although Laplacian blur detection is a good indicator is a good indicator of blurred images, the algorithm failed in this case. The failure of the blur detection algorithm can be attributed to poorly taken images and inconsistent image sizes. We tried to change the size of the images, however, it did not improve the accuracy of the results. Probably using new deep learning based methods will be more effective.

### 3.4 Privacy implications of VizWiz

In the analysis of VizWiz, we have seen various issues of people with visual impairment. Definitely, such applications are helping the users, making them more independent. However, there are privacy risks. We have seen people share their medical health information, often they share their address web workers. The authors of VizWiz data set did not share 5000 photos due to privacy reason. However, people often share their credit card information which can have severe consequences. The information given to unfamiliar people can be exploited. Therefore, additional care is required for such data. Based on the analysis, we have seen multiple times that it is not always possible to automatically answering the questions. We need human intelligence for some challenges. If the data requires human intelligence, then instead of sending the complete data the system can send partial data so that the privacy implication can be reduced.

Another potential privacy threat can be arose from the inability to know what is in the picture. The user can mistakenly capture sensitive photos and share it with the web workers. The bystanders of such devices are also in risk, because they can also inadvertently captured by the user and shared with the crowd workers. One such example is shown in Figure 13. If we check the figure, we can see that a bystander is present in the picture. The question asked for this question was ‘What is this?’. We can assume that the user probably was trying to detect an object but took a photo of nearby person. Similar privacy leakage can happen with credit cards, and other sensitive information. Photos can be shared in error. Therefore, such systems should consider such implications and should take extra precaution to reduce such incidents.

[Figure 13 about here.]

## 4 CONCLUSION

People with visual impairments faces different challenges and by analyzing the VizWiz data set we identified and explored some challenges. Although some challenges could be identified by analyzing portions of the data, big data analytics helps us to get a better exploration of the challenges. Moreover, big data analytics also helps to discover some solution space. In future, if other such services similar analysis it would be possible to reduce the human effort that is required to operate such services. Moreover, with more data

it would be possible to early detect the risks. By early detecting the risks, the system would be more helpful for people with visual impairments. Only in that way, they can enjoy the similar quality like other sighted people.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski for helping us with the instruction and resources that were required to complete this paper. We would also to like to thank the associate instructors for being available on the course website all the time and helping us with their answers.

## REFERENCES

- [1] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. 2016. Addressing Physical Safety, Security, and Privacy for People with Visual Impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 341–354. <https://www.usenix.org/conference/soups2016/technical-sessions/presentation/ahmed>
- [2] Aipoly. 2017. Vision through artificial intelligence. <http://aipoly.com/>. (2017).
- [3] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 333–342. <https://doi.org/10.1145/1866029.1866080>
- [4] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of CHI 2013*. <https://www.microsoft.com/en-us/research/publication/visual-challenges-in-the-everyday-lives-of-blind-people/>
- [5] Susumu Harada, Daisuke Sato, Dustin W. Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible Photo Album: Enhancing the Photo Sharing Experience for People with Visual Impairment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2127–2136. <https://doi.org/10.1145/2470654.2481292>
- [6] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting Blind Photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '11)*. ACM, New York, NY, USA, 203–210. <https://doi.org/10.1145/2049536.2049573>
- [7] Microsoft. 2017. Seeing AI: Turning the visual world into an audible experience. <https://www.microsoft.com/en-us/seeing-ai/>. (2017).
- [8] Michelle Naranjo. 2016 (accessed Sep 1, 2017). Toyota’s Project BLAID Is an Empowering Mobility Device for the Visually Impaired. <https://www.consumerreports.org/car-safety/toyota-project-blaid/>. (2016 (accessed Sep 1, 2017)).
- [9] Looktel Recognizer. 2017. Instantly recognize everyday objects. <http://www.loktel.com/recognizer>. (2017).
- [10] Adrian Rosebrock. 2015. Blur detection with OpenCV. <https://www.pyimagesearch.com/2015/09/07/blur-detection-with-opencv/>. (2015).
- [11] VizWiz. 2017. VizWiz DataSet. <http://www.vizwiz.org/data/>. (2017).

## A APPENDIX

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

#### LIST OF FIGURES

1	Seeing AI providing various information about people nearby [7].	8
2	Distribution of the number of responses	9
3	Most frequently used words in the questions asked	10
4	Most frequently used interesting words	11
5	Object Detection Questions	11
6	Color Detection images	12
7	Screen Reading images	12
8	Reading problems related images	13
9	Questions asked containing ‘look’ and ‘like’	13
10	Various health related questions	14
11	Poorly captured images	14
12	Accuracy of Laplacian blur detection	15
13	Privacy Implications of VizWiz	16
14	figure	17
15	figure	17
16	Most frequently used interesting words	18

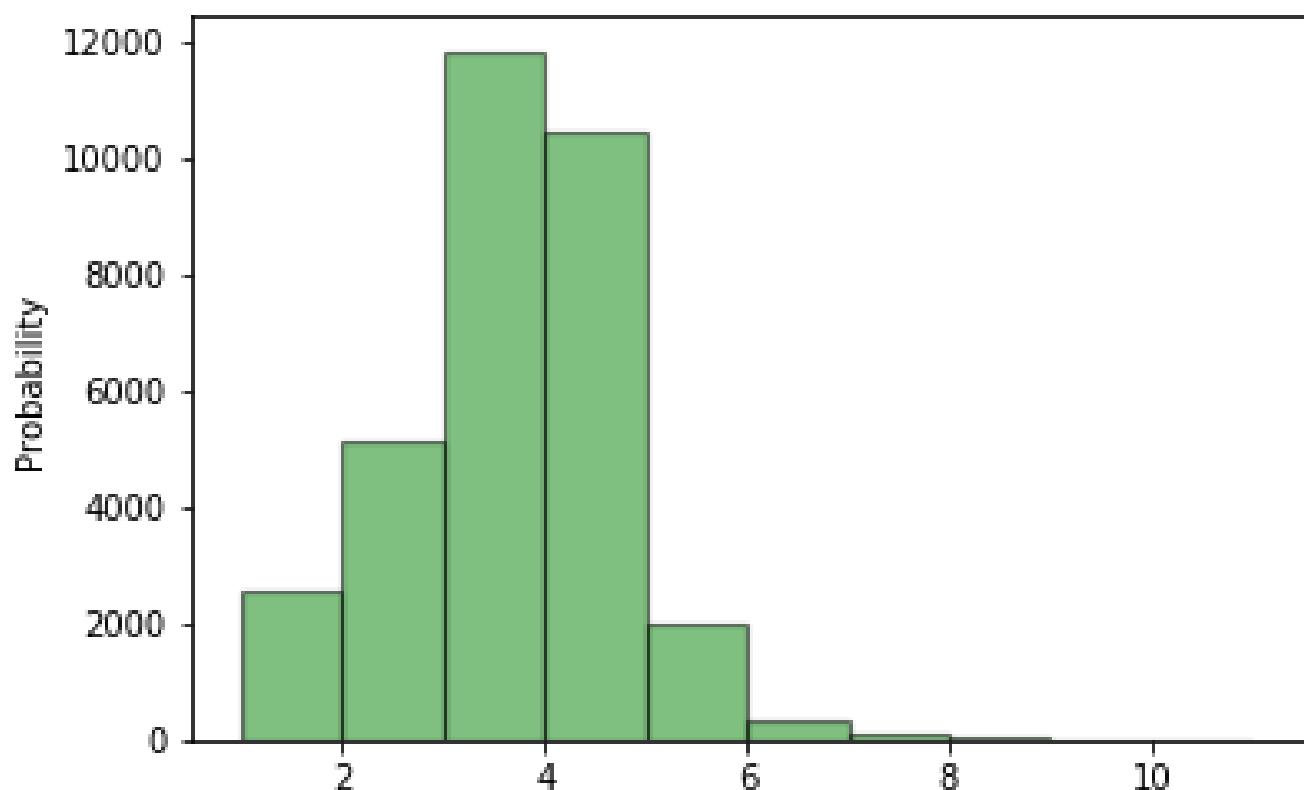


(a) Age, gender, appearance, and facial expression



(b) Age, gender, and activity of a nearby person

Figure 1: Seeing AI providing various information about people nearby [7].



**Figure 2: Distribution of the number of responses**

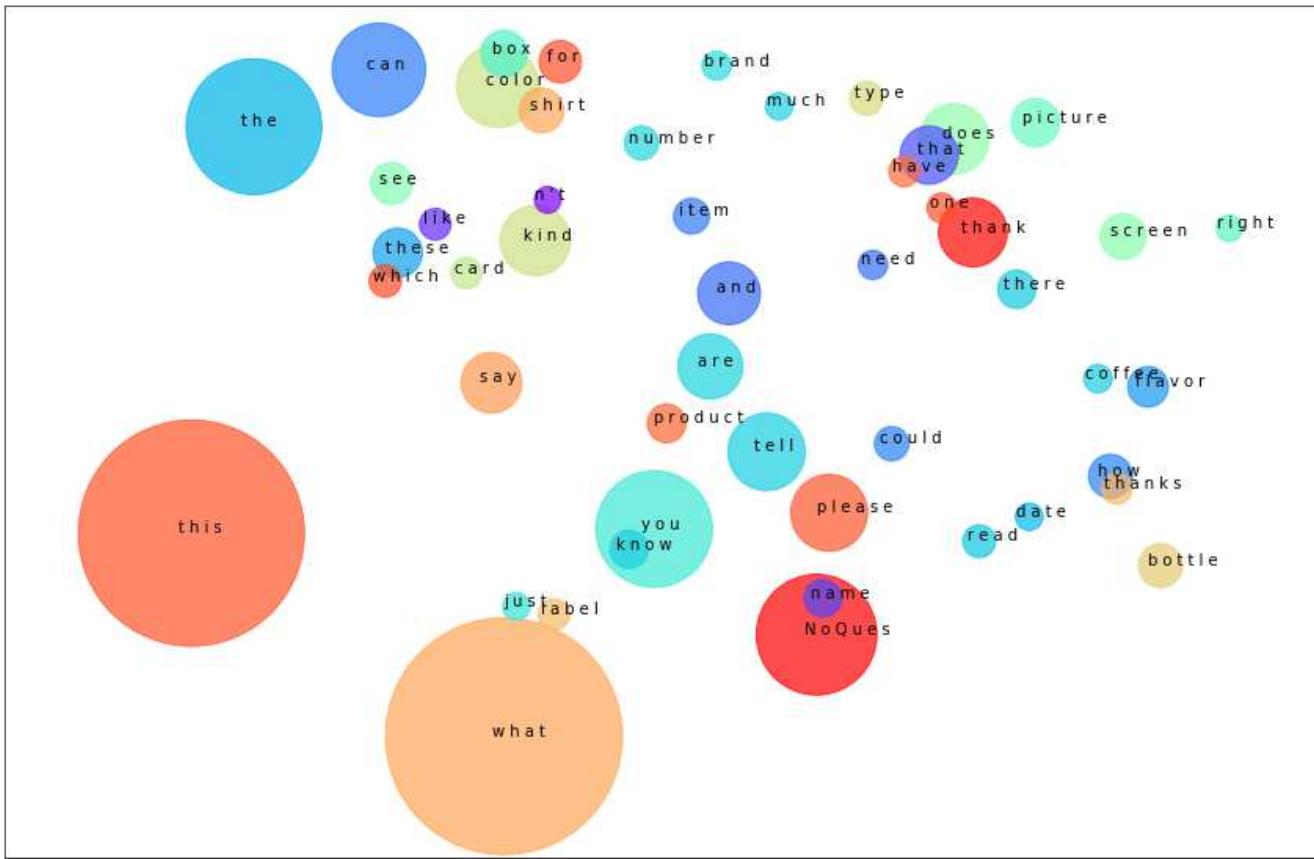


Figure 3: Most frequently used words in the questions asked

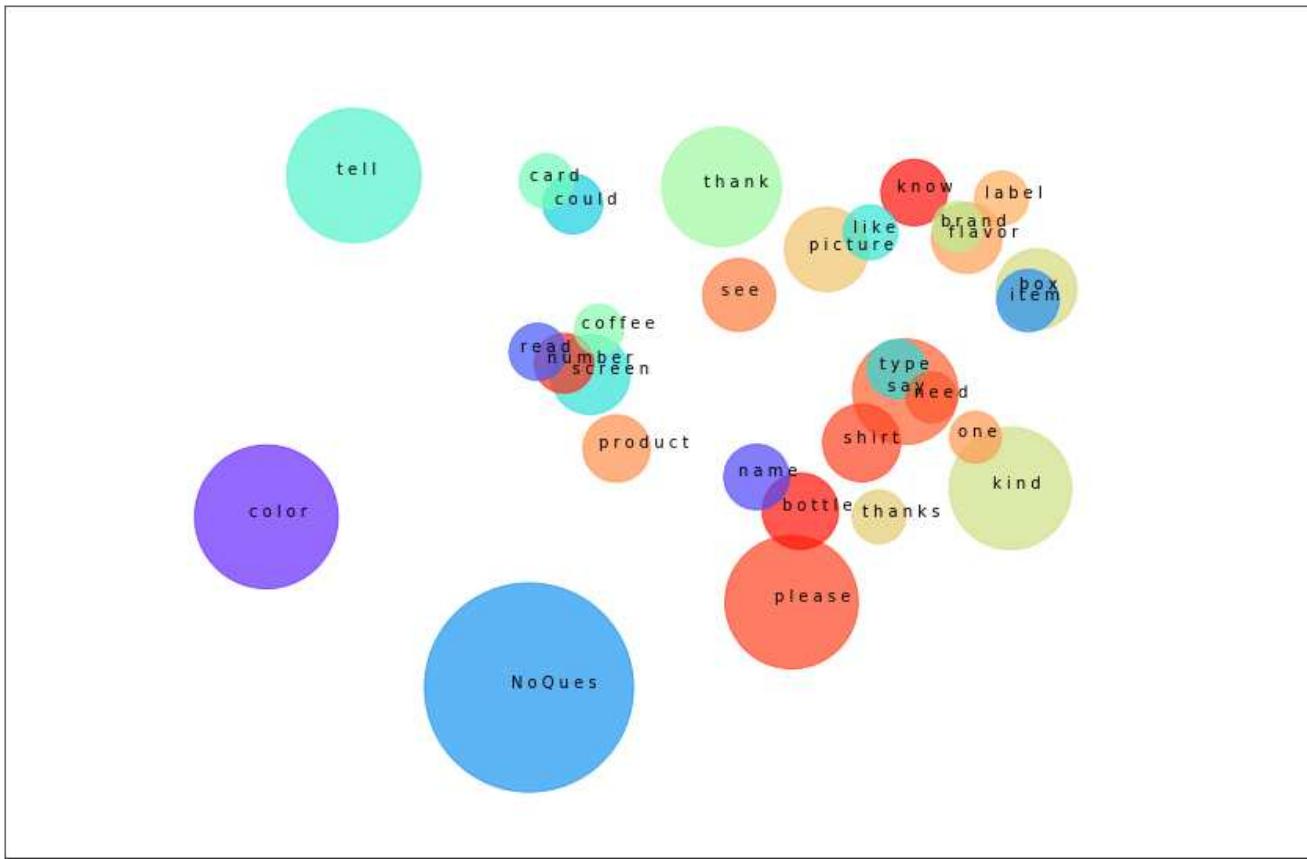


Figure 4: Most frequently used interesting words



Figure 5: Object Detection Questions

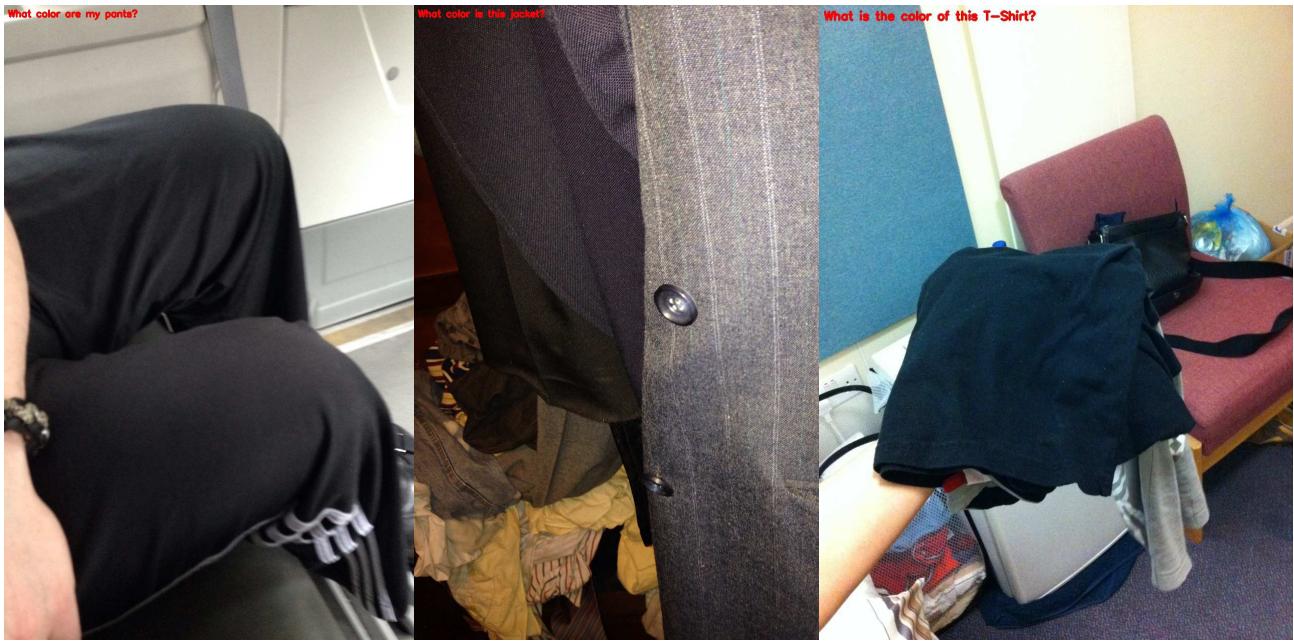


Figure 6: Color Detection images

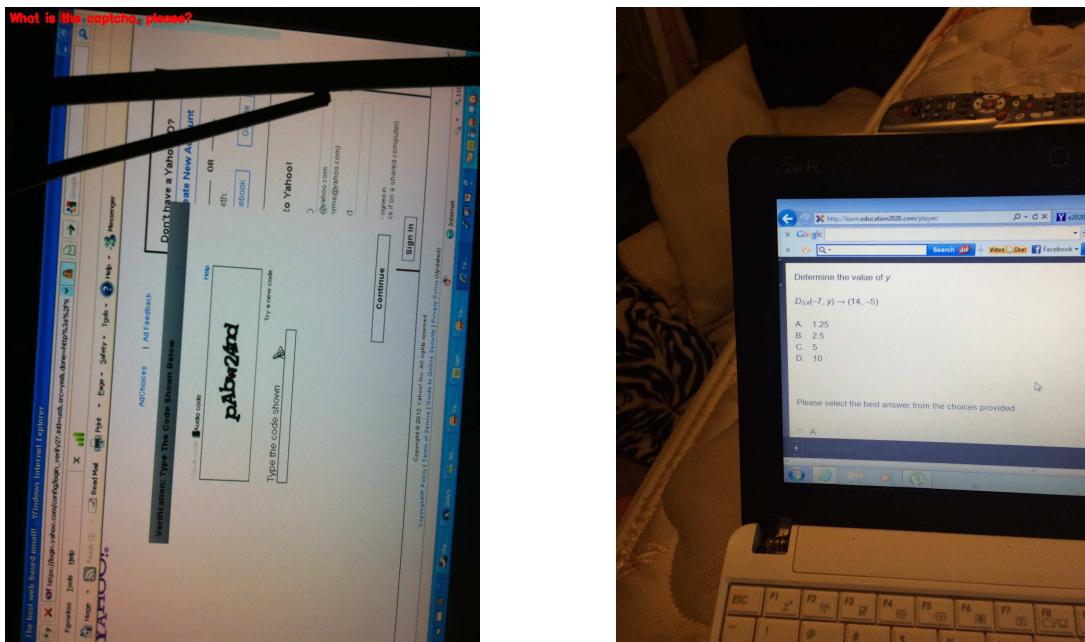


Figure 7: Screen Reading images



Figure 8: Reading problems related images

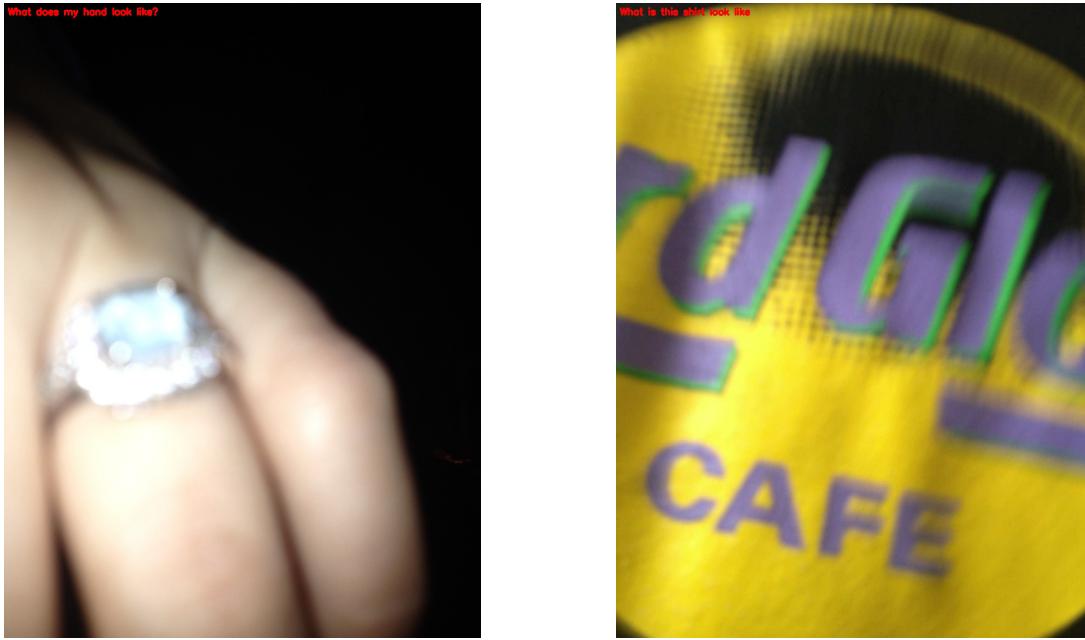
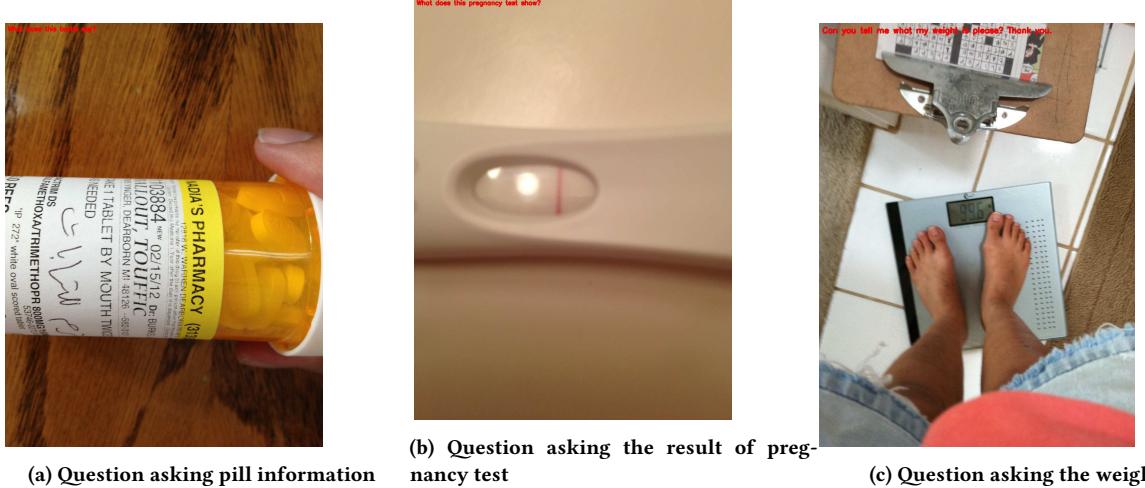


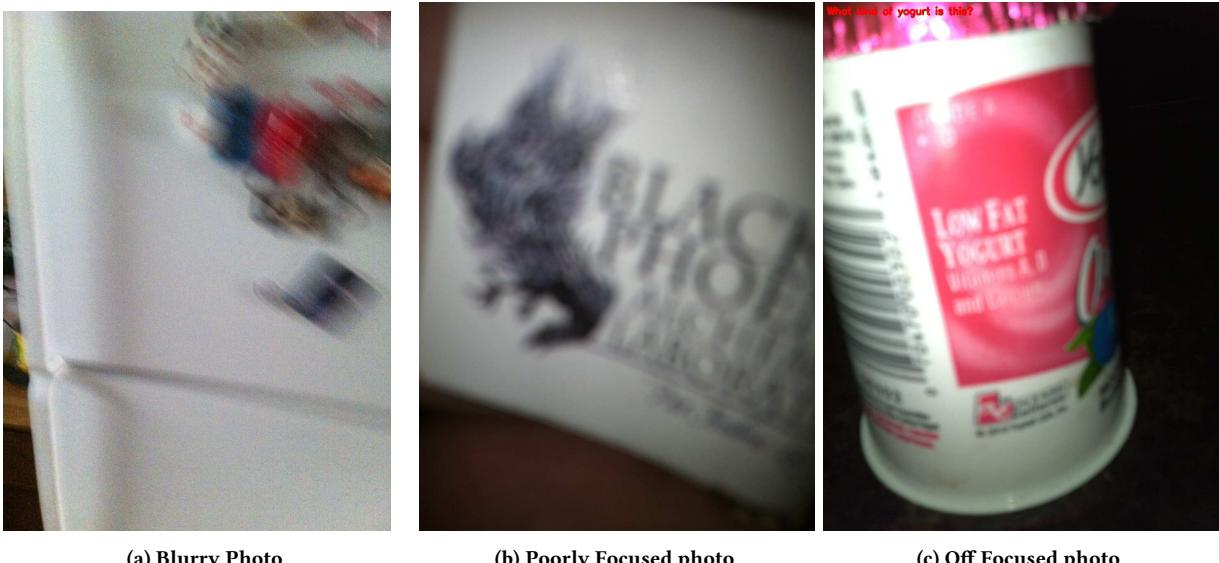
Figure 9: Questions asked containing 'look' and 'like'



(b) Question asking the result of pregnancy test

(c) Question asking the weight

Figure 10: Various health related questions

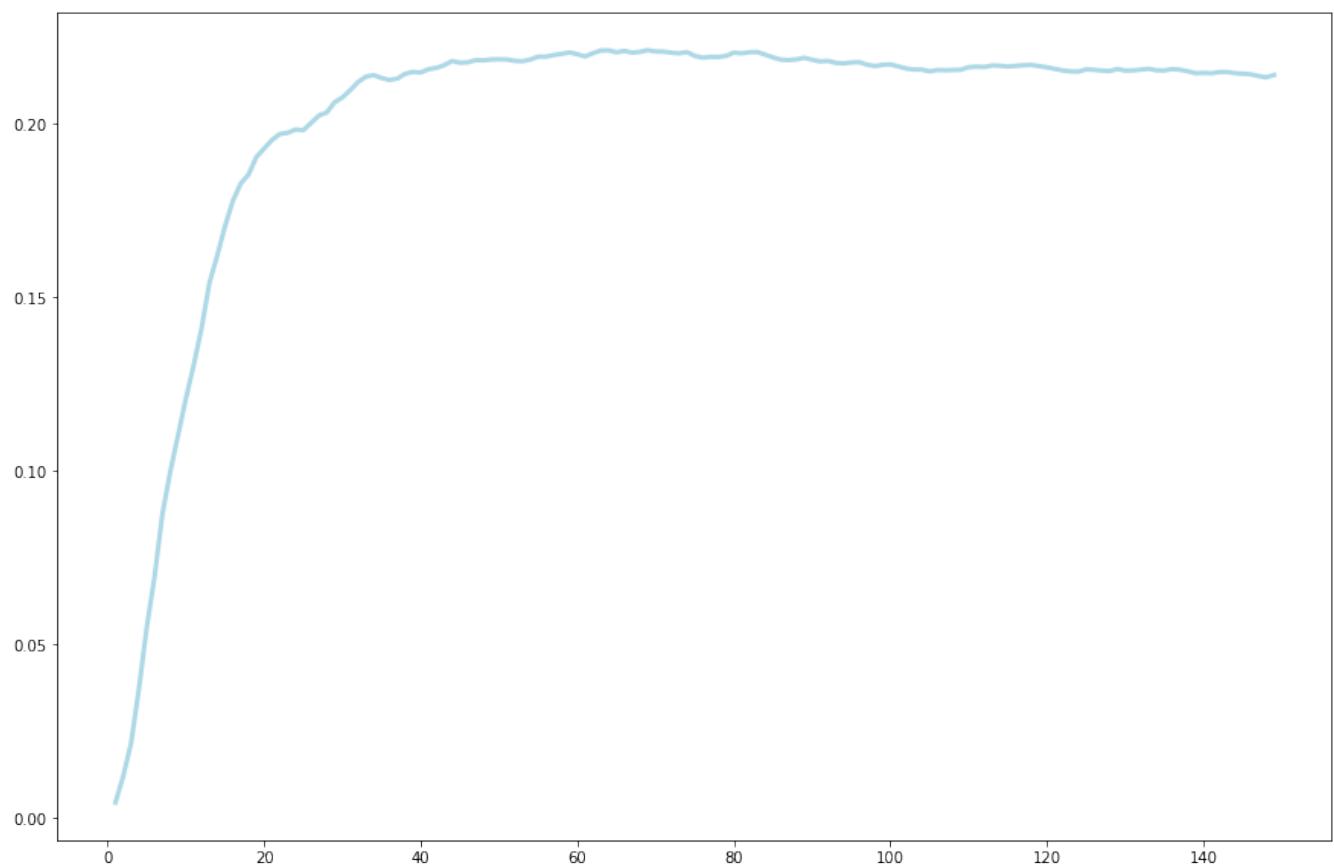


(a) Blurry Photo

(b) Poorly Focused photo

(c) Off Focused photo

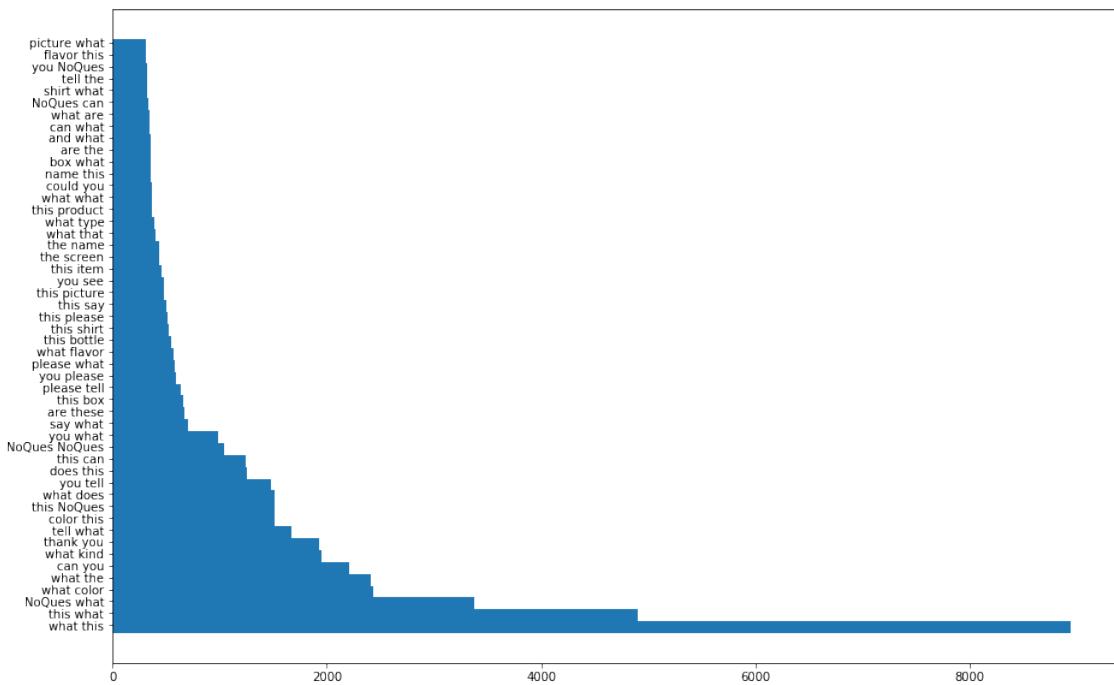
Figure 11: Poorly captured images



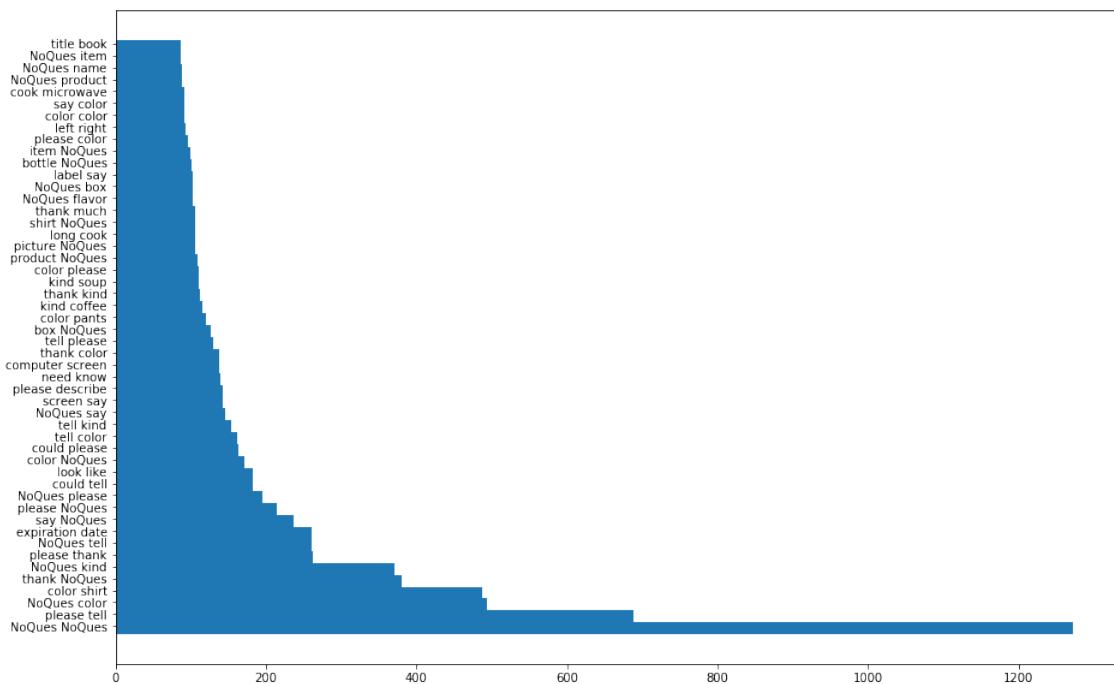
**Figure 12: Accuracy of Laplacian blur detection**



Figure 13: Privacy Implications of VizWiz  
404



**Figure 14: figure**  
Most frequently used pair of word



**Figure 15: figure**  
Most frequently used pair of interesting words

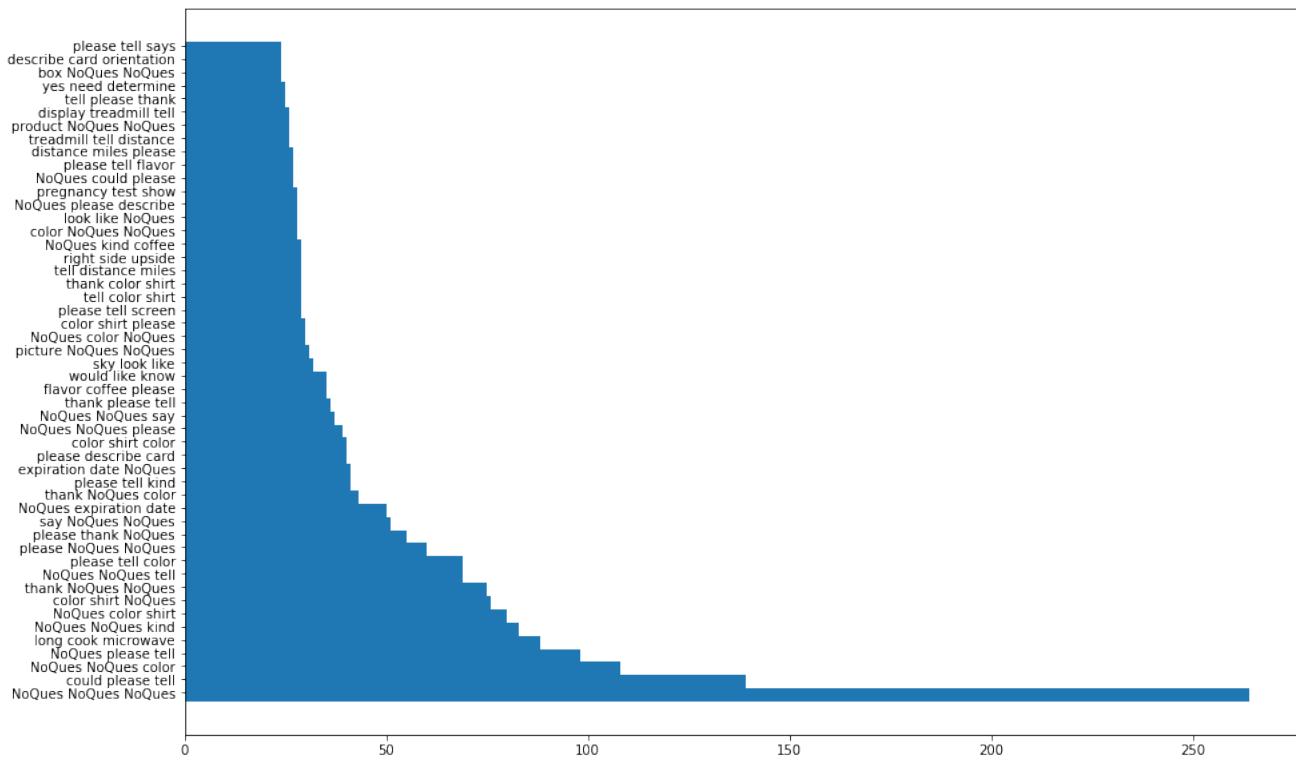


Figure 16: Most frequently used interesting words

```
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--numpages field, but no articleno or eid field, in Bigham:2010
Warning--no journal in Brady:2013
Warning--no number and no volume in Brady:2013
Warning--numpages field, but no articleno or eid field, in Harada:2013
Warning--numpages field, but no articleno or eid field, in Jayant:2011
(There were 5 warnings)
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
report.bib:108:@inproceedings {Ahmed:2016,
```

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-12-10 13.50.04] 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 51.24545pt.
Typesetting of "report.tex" completed in 1.7s.
```

---

```
Compliance Report
```

---

```
name: Ahmed, Tousif
hid: 237
paper1: 100%, October 27, 2017
```

```
paper2: 100%, Nov 6, 2017  
project: 100%, Dec 7, 2017
```

```
yamlcheck
```

---

```
wordcount
```

---

```
18  
wc 237 project 18 4961 report.tex  
wc 237 project 18 4849 report.pdf  
wc 237 project 18 598 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

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

---

```
passed: False
```

```
floats
```

---

```
43: Big data analytics could be a potential alternative. To understand  
how big data can help people with visual impairments, we need to  
understand the background first. Nowadays, people with visual  
impairments use different technologies for their problems. A wide  
range of technologies such as talking watch, braille reader,  
navigation helper are available in the market to help the visually
```

impaired in their daily tasks. Since the introduction of smartphone, smartphone based applications gained huge popularities among people with visual impairments. Now, mobile and smartphone applications like Seeing AI<sup>[1]</sup>, AiPoly<sup>[2]</sup>, LookTel<sup>[3]</sup>, and other such camera based applications are helping people with visual impairments in object recognition, face recognition, color detection, human emotion detection, activity recognition, and other such tasks that was not possible before. Figure <sup>[4]</sup> depicts an example from Seeing AI which shows that how camera based applications are helping people with visual impairments by describing nearby person's activity (Figure <sup>[5]</sup>) and their facial emotions (Figure <sup>[6]</sup>) .

```

44: \begin{figure}[tbp]
47: \includegraphics[scale=0.3]{images/facial.pdf}
49: \label{fig:facial}
53: \includegraphics[scale=0.3]{images/activity2.pdf}
55: \label{fig:ios}
59: \label{fig:seeingai}
83: \item \textbf{response}: Each question can have multiple responses. As mentioned earlier, some questions were tried to answer using the IQ engine and some questions were sent to the web workers. For each question, there can be one to 11 responses. However, on average three responses were received. The distribution of the responses shown in Figure[7]. From the figure, we can see that most of the questions either received three or four responses.
84: \begin{figure}![ht]
85: \centering\includegraphics[width=\columnwidth]{images/r_count.png}
87: \label{f:response count}
105: \begin{figure}[hbp]
108: \includegraphics[width=\columnwidth]{images/uni_all.png}
110: \label{fig:uni_all}
114: To understand the challenges, we first calculated the frequency of the words. There are around 4500 unique words in the questions. The most frequent 50 words is shown in Figure[8]. If we closely examine the words, we can see that the most frequently used word is \emph{'what'}. \emph{'What'} appeared 22793 times which is approximately 70\% of all of the worlds. The second and third most frequent words are 'this' and 'the'. Since, this is a set of questions, therefore, all the above words are justifiable. Although, 'what' is somewhat giving us an indication that users are asking about objects or subjective questions mostly, 'this and 'the' is not adding that much value. Next, we performed the same analysis by removing the most commonly used words in English. That unigrams

```

gave us some additional insights. The list of most frequently used interesting words can be found in Figure `\ref{fig:uni_int}`. If we remove the commonly used words, then for the majority of the questions had no questions. Those questions were asked by just uploading the photos. We assume that the users thought that the app could automatically answer those questions. Other three most frequently used words are 'color', 'tell', and 'please'. Among these three the most interesting is 'color'. Combination of 'what' and 'color' indicates that people with visual impairments faces issues with color detection, and often they ask the workers about the color of the objects and items. Therefore, we found `\textbf{color detection}` problem of people with visual impairments from the analysis. If we just consider the nouns and pronouns from the 30 most frequently used words, we find 'box', 'picture', 'color', 'screen', 'shirt', 'bottle', 'flavor', 'brand', 'coffee', 'label', and 'product'. From this keywords, we can safely assume three other problems: they face issues with screens (screen), there are issues with objects (brand), and the users face issues with reading labels. Therefore, from the initial analysis we found four problems that people with visual impairments regularly face: `\textbf{color detection}`, `\textbf{object detection}`, `\textbf{reading screens (mobile/computer)}`, and `\textbf{reading labels}`.

```

116: \begin{figure}[hbp]
119: \includegraphics[width=\columnwidth]{images/int_count.png}
121: \label{fig:uni_int}
126: After checking the most frequently used words, we explored the most interesting pairs of words. If we check the bigrams (Figure \ref{fig:bi_all} and \ref{fig:bi_int} in Appendix), it gives confidence of our identified problems. The most frequently used two words are 'what' and 'this' which suggests that most questions were asked to identify the object. Therefore, people with visual impairments definitely face problems with detecting objects. 'What' and 'color' also suggests that users face color detection problem frequently. If we check the bigrams of most frequently used interesting words (Figure \ref{fig:bi_int}), we find some additional insights. If we ignore 'NoQues', then we again see color detection and computer screen reading problem. However, now we can find another interesting pairs of words 'look' and 'like'. This pair indicates a subjective question, where the user is asking how the user is looking like. This identifies another challenges of people with visual impairments \textbf{Impression Management}. Another interesting common pair of words are 'long' and 'cook' which indicates reading label issues, however this can be a household activity issue. The trigrams also gave us some new interesting insights (Figure
  
```

`~\ref{fig:tri_int}).` Most of the trigrams confirms above mentioned challenges, however, there are some new issues. One interesting trigram is ‘display’, ‘treadmill’, and ‘tell’ which indicates the health fitness related issues or `\textbf{Health Management}` Issue. Due to the accessibility issues in health monitoring and fitness monitoring issues, they can not manage health effective. Therefore, the users often seek help for reading the display. Another interesting three words are ‘pregnancy’, ‘test’, ‘show’ which can also be put into health Management category. However, this seems a private information, but still people with visual impairments have to share this information due to their visual challenges.

- 132: `\subsubsection{Object Detection:}` The most frequently asked question in VizWiz is ‘What is this’ or ‘What is that’. ‘What’ appeared more than 22,000 questions. Among those 22,000 questions around 7,000 questions are ‘What is this?’ and ‘What is that?’. People asks variety of object detection questions ranging from everyday objects to personal objects. Some examples of object detection problem is shown in Figure `~\ref{fig:object}`. By manually analyzing some photos, it seems most of them are related to household activities. Therefore, with better tools it is possible to detect the objects.
- 133: `\begin{figure}[hbp]`
- 136: `\includegraphics[scale=0.3]{images/object_1.pdf}`
- 140: `\includegraphics[scale=0.3]{images/object_2.pdf}`
- 144: `\label{fig:object}`
- 149: `\subsubsection{Color Detection:}` Another most frequent problem that people with visual impairments face is to detect colors. Most of the time they use VizWiz to identify colors of their cloths, items, foods, and others. Some examples of color detection is shown in Figure `~\ref{fig:color}`. Based on the images, automatically detecting the colors seems a challenging task. Because, if we examine figure `~\ref{fig:color}` we can see in the image there can be other objects. Automatically detecting the object of interest will be difficult. For example, in the right most photo the user is asking about the color of the dress in hand, however, there are other objects visible in the photo. Therefore, identifying the color automatically will be challenging.
- 150: `\begin{figure}[hbp]`
- 153: `\includegraphics[scale=0.3]{images/color_1.pdf}`
- 157: `\includegraphics[scale=0.3]{images/color_2.pdf}`
- 160: `\includegraphics[scale=0.3]{images/color_3.pdf}`
- 164: `\label{fig:color}`
- 167: `\subsubsection{Reading Screens:}` Nowadays, people with visual impairments use smartphones and computers. They use screen

reading software which generates synthesized speech to relay the information from screen. However, some times these software fail and visually impaired need to seek help from crowd workers.

Another issue is the accessibility issues of CAPTCHA, people with visual impairments struggles with CAPTCHA. Therefore, they seek people who can read the CAPTCHA for them. Some examples of reading screen problems are shown in Figure~\ref{fig:screen}.

```
168: \begin{figure}[hbp]
171: \includegraphics[scale=0.3]{images/screen_1.pdf}
175: \includegraphics[scale=0.3]{images/screen_2.pdf}
180: \label{fig:screen}
183: \subsubsection{Reading documents or labels:} Another obvious challenges of people with visual impairments is reading documents. The paper documents are not often accessible and people need help from others to read that. People might use scanners to read documents, however, scanning documents can be time consuming. Especially, for scanning food or medicine labels can be difficult. Therefore, participants seek help to read labels for them. Figure~\ref{fig:reading} shows some examples of reading issues. However, there can be potential score for technology for this types of problem. If the user is asking for reading helps, a simple OCR can help. However, OCR might not work well with food labels. One suggestion could be for food related reading question, the system could look for barcode and identify necessary information.
184: \begin{figure}[hbp]
187: \includegraphics[scale=0.3]{images/reading_1.pdf}
191: \includegraphics[scale=0.3]{images/reading_2.pdf}
194: \includegraphics[scale=0.3]{images/reading_3.pdf}
199: \label{fig:reading}
202: \subsubsection{Impression Management:} Based on the analysis, we explored that managing impressions can be challenging. As a social norm, we often present our better selves to others by wearing consistent dresses. For example, we do not want to present ourselves in social places in such way that may misrepresent ourselves. Some words that we found in the questions are ‘look’, ‘like’ which we assume that users are asking to understand their appearance. Therefore, impression management for people could be challenging. Sometimes, the questions can be appearance related. Some examples of impression management challenges is shown in Figure ~\ref{fig:impression}.
203: \begin{figure}[hbp]
206: \includegraphics[scale=0.3]{images/impression_1.pdf}
210: \includegraphics[scale=0.3]{images/impression_2.pdf}
215: \label{fig:impression}
218: \subsubsection{Health Management:} Health management is important
```

for everyone. However, people with visual impairments face lot of challenges to maintain healthy behavior. They struggles to cook, therefore, they need to eat outside or eat packaged foods. They can not read the package's well, so miss the nutrition info.

Managing medicine can be issue. Some other issues can be attributed to visual representation of results. For examples, weight scales show visual weights, pregnancy scales convey visual feedback, health monitoring instruments like treadmill convey visual information. All these visual information makes it difficult for managing health issues. Therefore, health management can be challenging. For that reasons, people with visual impairments often ask such applications to help them with various visual indicators in health and fitness. Figure `\ref{fig:health}` shows three different health realted issues of people with visual impairments. Figure `\ref{fig:pill}` depicts the issues of medicine management, users often can not identify the required medicine. Figure `\ref{fig:preg}` shows asking the result of pregnancy test, which can be sensitive.

Figure `\ref{fig:weight}` asking questions about the weight of the user. Since, such applications can forward these questions to friends and family members all these images can be sensitive. However, technology can potentially address this issue by automating the responses.

```
219: \begin{figure}[hbp]
222: \includegraphics[scale=0.2]{images/health_1.pdf}
224: \label{fig:pill}
228: \includegraphics[scale=0.2]{images/health_2.pdf}
230: \label{fig:preg}
233: \includegraphics[scale=0.2]{images/health_3.pdf}
235: \label{fig:weight}
239: \label{fig:health}
244: \begin{figure}[hbp]
247: \includegraphics[scale=0.15]{images/blurry.jpg}
249: \label{fig:blur}
253: \includegraphics[scale=0.25]{images/poor_f.pdf}
255: \label{fig:pf}
258: \includegraphics[scale=0.25]{images/off_focus.pdf}
260: \label{fig:off}
264: \label{fig:photo}
266: Figure \ref{fig:photo} depicts the some not understandable photos taken by people with visual impairments. However, such images takes resources and often cost money. If the system can early detect such images and prevent those images from sending then it can save resources. Misplaced or blurry photos can be early detected. Another potential scope of technology is to automatically fix the blurry images.
```

```

281: \subsubsection{Identifying a good threshold} We run the algorithm
      with various thresholds. The F1 score graph against various
      thresholds did not improve the accuracy. Figure
      \ref{fig:accuracy} shows the accuracy of blur detection.
282: \begin{figure}[hbp]
285: \includegraphics[width=\columnwidth]{images/f1.png}
287: \label{fig:accuracy}
297: Another potential privacy threat can be arose from the inability
      to know what is in the picture. The user can mistakenly capture
      sensitive photos and share it with the web workers. The
      bystanders of such devices are also in risk, because they can
      also inadvertently captured by the user and shared with the crowd
      workers. One such example is shown in Figure \ref{fig:privacy}. If
      we check the figure, we can see that a bystander is present in
      the picture. The question asked for this question was ‘What is
      this?’. We can assume that the user probably was trying to detect
      an object but took a photo of nearby person. Similar privacy
      leakage can happen with credit cards, and other sensitive
      information. Photos can be shared in error. Therefore, such
      systems should consider such implications and should take extra
      precaution to reduce such incidents.
298: \begin{figure}[tbp]
301: \includegraphics[width=\columnwidth]{images/private_1.pdf}
303: \label{fig:privacy}
327: \includegraphics[scale=0.45]{images/bigram_all.png}
329: \label{fig:bi_all}
336: \includegraphics[scale=0.45]{images/bigram_interesting.png}
338: \label{fig:bi_int}
346: \includegraphics[scale=0.5]{images/trigram_interesting.png}
348: \label{fig:tri_int}

```

```

figures 13
tables 0
\includegraphics 28
labels 24
refs 13
floats 13

```

```

True : ref check passed: (refs >= figures + tables)
False : label check passed: (refs >= figures + tables)
False : 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--numpages field, but no articleno or eid field, in Bigham:2010  
Warning--no journal in Brady:2013  
Warning--no number and no volume in Brady:2013  
Warning--numpages field, but no articleno or eid field, in Harada:2013  
Warning--numpages field, but no articleno or eid field, in Jayant:2011  
(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 in Genomics and Medicine

Matthew Durbin, MD FAAP

Indiana University School of Medicine Department of Pediatrics

Division of Neonatology Riley Hospital for Children

699 Riley Hospital Drive

Indianapolis, Indiana 46202

mddurbin@iu.edu

## ABSTRACT

The entirety of the human genome was sequenced in 2003, ushering in a new era of molecular biology, genetics and medicine. Since that time, technologies have advanced significantly, and next generation sequencing allows increasingly rapid and affordable sequencing of the entire human genome. Beyond the human genome, we can also sequence the entire RNA transcriptome, proteome, and metabolome. We can compare these entities in health and disease, and across populations. These new technologies produce massive datasets. Big Data applications and analytics allow interpretation and utilization of these data sets. However, analyzing and interpreting these datasets lags behind sequencing technology, as the rate limiting step. Still these technologies hold great potential for advancing medicine and human health. Combining this omics data with the electronic health record, wearable technology, pharmaceuticals and procedures, moves us towards personalized, precision, medicine.

## KEYWORDS

i523, hid311, Big Data, genomics, genetics, species reintroduction, environment, conservation

## 1 INTRODUCTION

### 1.1 Health and Genomics

The current state of healthcare system in the United States is often described as a crisis. The term comes with good reason, as spending accounts for 17-18% of GDP, dwarfing other nations, and is exponentially rising at an unsustainable rate. For all of our spending, we have poorer health than most developed and many developing nations. The healthcare industry is behind in technology, with recent adoption of an electronic medical record, and prior reliance on paper charting. Communication is most often by decades old technology including phone or fax. Internet communication between healthcare providers, and with patients, is a recent novelty. We have the poorest health, including obesity due to poor diet, lack of exercise, and substance abuse. We pay more for pharmaceuticals than any other country, and most pharmaceutical budget goes to marketing as opposed to research and development. To determine a familial or genetic risk for disease we mostly rely on patient interview.

Big Data has major potential to impact health. Massive data sets related to human health and genomics are compiled by insurance companies, pharmaceutical companies, public health institutions and research institutions. [6] Healthcare is making strides and big data collection is visible everywhere. The electronic medical record EMR is close to universal and is improving constantly. Medical

resources are accessible around the world through smartphones. Wearable technology and fitness tracking apps, nutrition apps are improving personal health. One of the biggest potential impacts to health comes with the advances in next generation sequencing and genomics. These new technologies allow us to determine genetic disease risk, determine prognosis, and predict response to pharmacology and other treatment, all by measuring the genetic code. The most powerful application will be to combine genomic data with the data generated from the EMR, wearable technology, model systems, etc. to develop personalized medicine strategies.

### 1.2 The Genetic Code

For centuries we have known much disease is heritable. Taking a thorough family health history via patient has been a mainstay of medical interview. Previously the medical provider merely noted ailments that ran in families and maintained vigilance in subsequent generations. All of that changed in 1953 when James Watson and Francis Crick reported the molecular structure of Deoxyribonucleic acid, or DNA [27]. DNA is a relatively simple structure is made up of four nucleotides, adenine, guanine, cytosine and uracil on a carbohydrate background. Different triplicate combinations of these four nucleotides, code for 20 amino acids, and these 20 amino acids make up every protein in all living things. This relative simple system is called the genetic code. Much like the 0's and 1's in computer code, giving rise to the complexity of the internet, the genetic code gives rise to the complexity of all living things. Each organism has a unique genetic code, and this molecular blueprint is utilized to create their protein, carbohydrate, lipid structure. Furthermore this DNA code is replicated, blended through reproduction, and passed to future generations. This genetic code is often interrupted in disease such as cancer. Differences in the genetic code lead to differences in disease susceptibility, and treatment response. The human genome refers to humans over 3 billion nucleotides. We have the ability to sequence the human genome, or determine the order of these nucleotide bases.

### 1.3 The Human Genome

The first human genome was sequenced in 2003 [3]. This colossal global effort took over 10 years and thousands of scientists working at great expense. In the end, a private and public group collectively sequenced the first genome. Initially, the technology was extremely expensive and took great deal of time. Through technological advancements including sequencing cores and big data, the cost of the genome has plummeted. The 1000-dollar genome project is an attempt to make sequencing more affordable [6]. We are a long way away from being able to utilize the genome to deliver care.

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 mandates 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
```

```
[2017-12-10 13.50.49] 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 0.9s.  
..../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
```

---

```
9  
wc 311 project 9 6590 report.tex  
wc 311 project 9 7166 report.pdf  
wc 311 project 9 1117 report.bib
```

```
find "
```

---

```
103: 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}
```

---

```
4: \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
```

# 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 [14]. 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 replaces 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 [21]. 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 [12].

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 files 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 [29].

### 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 [15].

## 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 [23]. 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 [27]. Goldman Sachs Global Investor Research group provides stock ratings on a scale of Buy, Neutral, and Sell [26].

J P Morgan Chase (JPM) is one of the largest investment banks in the world [28]. 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 [19].

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 [28]. 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 [31].

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 [22]. NASDAQ provides API services for subscribers starting at \$5,000 per year [20]. 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 example, the stock history for Celsion Corporation during the week of August 21, 2017 is shown:

```
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 [24]. The package can be installed through Git by cloning the Git directory where the package is available: [18]. To install the python package without Git the tape archive can be downloaded from [25]. Tape archives allow for compression of multiple files which can be restored to their original format using the tar command in the command line [16]. 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:

```
{"installed": {"client_id": "####.apps.googleusercontent.com",
  "project_id": "###",
  "auth_uri": "https://accounts.google.com/o/oauth2/auth",
  "token_uri": "https://accounts.google.com/o/oauth2/token",
  "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",
  "client_secret": "###",
  "redirect_uris": ["urn:ietf:wg:oauth:2.0:oob",
    "http://localhost"]}}
```

Once credentials are received in the form of a JSON file, the Google Client Library can be installed using pip to install google-api-python-client. The Google Development team has provided a quick-start 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 [13]. 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:googlealerts-noreply@google.com before:2017/11/24').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 about here.]

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 about here.]

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

Twelve out of sixty-three 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.

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

ABEO Snippet appears to reflect a number of disappointments followed by something positive:

Abeona Therapeutics – String Of 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 Celsion 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-deterrent opioid  
med The Pharma Letter Egalet (Nasdaq: EGLT)  
says its share move up a hefty 38.55%

This Google Alert was received on November 17, just prior to another 30% stock increase.

HCM is a pharmaceutical company headquartered in China. The Snippet reflects the company's 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 of analyst ratings:

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 The Finance  
Registrar The Company's lead program  
orthwest Bioth Cnn (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:

Otomy (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 [17]. 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 together 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 [30].

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 associated 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 about here.]

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 about here.]

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

### 3 RESULTS

The results export to a .csv as shown:

```
Accuracy , Date , Sentiment , Ticker  
L,2017-12-03,L,ABBV  
L,2017-12-01,L,ABBV  
L,2017-11-30,L,ACAD  
L,2017-12-02,L,ALNY  
L,2017-12-03,L,ARGX  
L,2017-12-02,L,BABA
```

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 sentiment ranking which is actually what would be considered a positive rating:

```
EPS for The Kroger Co. (KR) Expected At \$.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 does not confirm the model as the dataset was very low, 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.

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>
- [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] Thomas Hellstrom and Kenneth Holmstrom. 1997. *Predict the stock market*. techreport. Department of Mathematics and Physics, Mälardalen University, Sweden.
- [13] hugovk. 2017. Httpplib2. Website. (10 2017). <https://github.com/httpplib2/httpplib2>
- [14] Investopedia. 2017. Stock Market. Website. (09 2017). <https://www.investopedia.com/terms/s/stockmarket.asp?lgl=rira-layout>
- [15] 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>
- [16] LINFO. 2006. The tar Command. website. (07 2006). <http://www.linfo.org/tar.html>
- [17] Steven Loria. 2017. TextBlob. Website. (01 2017). <http://textblob.readthedocs.io/en/dev/quickstart.html>
- [18] Lukaszbanasiak. 2016. Yahoo-finance. Website. (12 2016). <https://github.com/lukaszbanasiak/yahoo-finance>
- [19] 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>
- [20] NASDAQ. 2017. NASDAQ DataOnDemand Subscription Plans. Website. (2017). <https://www.nasdaqdod.com/Shop/ProductConfig.aspx?product=webservices&service=NASDAQDataOnDemand>
- [21] NASDAQ. 2017. NASDAQ's Story. website. (2017). <http://business.nasdaq.com/discover/nasdaq-story/index.html>
- [22] NASDAQ. 2017. U.S. Stock Quotae, Charts, and Research. website. (2017). <http://www.nasdaq.com/quotes/>
- [23] S.J. Pocock. 2013. *Clinical Trials: A Practical Approach*. Wiley, England. <https://books.google.com/books?id=TxbTBQAAQBAJ>

- [24] Python.org. 2016. yahoo-finance 1.4.0. Website. (11 2016). <https://pypi.python.org/pypi/yahoo-finance>
- [25] Python.org. 2016. Yahoo-finance 1.4.0. Website. (11 2016). <https://pypi.python.org/pypi/yahoo-finance>
- [26] Goldman Sachs Global Investment Research. 2017. Equity Ratings. Website. (01 2017). [http://www.goldmansachs.com/research/equity\\_ratings.html](http://www.goldmansachs.com/research/equity_ratings.html)
- [27] Goldman Sachs. 2017. At A Glance. Website. (01 2017). <http://www.goldmansachs.com/who-we-are/at-a-glance/index.html>
- [28] Shobhit Seth. 2013. The World's Top 10 Investment Banks. Website. (2013). <https://www.investopedia.com/articles/investing/111114/worlds-top-10-investment-banks.asp>
- [29] Chicago Tribune. 2008. Internet-fueled panic sinks stock. Website. (09 2008). [http://articles.chicagotribune.com/2008-09-09/news/0809090607\\_1-united-airlines-united-stock-bloomberg](http://articles.chicagotribune.com/2008-09-09/news/0809090607_1-united-airlines-united-stock-bloomberg)
- [30] Wendykan. 2017. DeepLearningMovies. Website. (03 2017). <https://github.com/wendykan/DeepLearningMovies>
- [31] Wiki. 2017. Timeline of machine learning. Website. (09 2017). [https://en.wikipedia.org/wiki/Timeline\\_of\\_machine\\_learning](https://en.wikipedia.org/wiki/Timeline_of_machine_learning)

## LIST OF FIGURES

- 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. 11
- 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. 12
- 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 13
- 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. 14



```

\begin{verbatim}
```
Tiffany Fabianac Modified code from:
Reading GMAIL using Python
  - https://github.com/abhishekchhibber/Gmail-Api-through-Python
- Abhishek Chhibber

This script does the following:
- Go to Gmail inbox
- Find and read all the Google Alert messages
- Extract details (Date, Snippet) and export them to a .csv file / DB

Before running this script, the user should get the authentication by following
the link: https://developers.google.com/gmail/api/quickstart/python
Also, client_secret.json should be saved in the same directory as this file
```

# Importing required libraries
from apiclient import discovery
from apiclient import errors
from httplib2 import Http
from oauth2client import file, client, tools
import base64
from bs4 import BeautifulSoup
import re
import time
import dateutil.parser as parser
from datetime import datetime
import datetime
import csv
import json
import io

# Creating a storage.JSON file with authentication details
SCOPES = 'https://www.googleapis.com/auth/gmail.modify' # we are using modify and not readonly, as we
# will be marking the messages Read
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'

# Getting all the unread messages from Inbox
# labelIds can be changed accordingly
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':

```

```

\begin{verbatim}
...
Collect Historical Stock Data
Tiffany Fabianac Modified code from:
  - http://pandas-datareader.readthedocs.io/en/latest/remote_data.html
...
from pandas_datareader import data
import pandas as pd
import csv
import string
import datetime
from collections import defaultdict
from pandas.tseries.offsets import BDay

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

with open('google_alert_data.csv' , 'rb') as csvfile:
with open('labeledTrainData.csv','wb') as f:
datareader = csv.DictReader(csvfile)
writer = csv.DictWriter(f, fieldnames=datareader.fieldnames , extrasaction='ignore' , delimiter=',' ,
skipinitialspace=True)
writer.writeheader()
for row in datareader:
writer.writerow(row)

```

```

\begin{verbatim}
...
Use KaggleWord to produce random forest analysis
Tiffany Fabianac Modified code from:
- https://youtu.be/AJVP96tAWxw
- Siraj Raval
...

import os
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.ensemble import RandomForestClassifier
from KaggleWord2VecUtility import KaggleWord2VecUtility
import pandas as pd
import nltk

if __name__ == '__main__':
    train = pd.read_csv(os.path.join(os.path.dirname(__file__), 'labeledTrainData.csv'), header=0,
                        delimiter=",", quoting=3)
    test = pd.read_csv(os.path.join(os.path.dirname(__file__), 'testData.csv'), header=0, delimiter=",",
                       quoting=3)

    print 'The first review is:'
    print train['Snippit'][0]
    raw_input("Press Enter to continue...")

    print 'Download text data sets'
    nltk.download()
    clean_train_reviews = []
    print "Cleaning and parsing the training set...\n"
    for i in xrange(0, len(train['Snippit'])):
        clean_train_reviews.append(" ".join(KaggleWord2VecUtility.review_to_wordlist(train['Snippit'][i],
   True)))

    print "Creating the bag of words...\n"
    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()

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

    print "Cleaning and parsing \n"
    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)
    output = pd.DataFrame(data={"Accuracy": "", "Sentiment": result, "Ticker": test["Ticker"], "Date": test["Date"]})
    output.to_csv(os.path.join(os.path.dirname(__file__), 'randomForestResults.csv'), index=False,
                  quoting=3)
    print "Wrote results to randomForestResults.csv"
\end{verbatim}

```

**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**

```

\begin{verbatim}
...
Validate random forest analysis
Tiffany Fabianac Modified code from:
  - http://pandas-datareader.readthedocs.io/en/latest/remote_data.html
...
from pandas_datareader import data
import pandas as pd
import csv
import string
import datetime
from collections import defaultdict
from pandas.tseries.offsets import BDay

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
endDate = datetime.datetime.today().strftime('%Y-%m-%d')

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

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

result = stockData(startDate , endDate , ticker)
if(result.iloc[0][ 'high' ] != result.iloc[0][ 'high' ]):
return 1
else :
open = result.iloc[0][ 'open' ]
return open

```

## 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
```

bibtex \_ label error

bibtext space label error

bibtext comma label error

latex report

```
50:20      error      trailing spaces  (trailing-spaces)
71:81      error      line too long (83 > 80 characters)  (line-length)
```

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

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

```
yamlcheck
```

---

```
wordcount
```

---

```
14
wc 313 project 14 7503 report.tex
wc 313 project 14 14903 report.pdf
wc 313 project 14 982 report.bib
```

```
find "
```

---

```
96: {"installed":{"client_id": "###.apps.googleusercontent.com",
97: "project_id": "###",
98: "auth_uri": "https://accounts.google.com/o/oauth2/auth",
99: "token_uri": "https://accounts.google.com/o/oauth2/token",
100: "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/
v1/certs",
101: "client_secret": "###",
102: "redirect_uris": ["urn:ietf:wg:oauth:2.0:oob",
103: "http://localhost"]}}
```

```
202: with open("API_out.csv", "a") as f:  
203:     header=["Date", "Snippet"]  
226: $ PATH="$PATH:/c/Python27"  
  
289: temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")  
  
298: temp_date = datetime.datetime.strptime(startDate, "%Y-%m-%d")  
  
460: train = pd.read_csv(os.path.join(os.path.dirname(__file__),  
        'labeledTrainData.csv'), header=0, delimiter=",", quoting=3)  
  
461: test = pd.read_csv(os.path.join(os.path.dirname(__file__),  
        'testData.csv'), header=0, delimiter=",", quoting=3)  
  
465: raw_input("Press Enter to continue...")  
  
470: print "Cleaning and parsing the training set...\n"  
  
472: clean_train_reviews.append(" ".join(KaggleWord2VecUtility.review_  
        to_wordlist(train['Snippet'][i], True)))  
  
474: print "Creating the bag of words...\n"  
  
475: vectorizer = CountVectorizer(analyzer="word", tokenizer=None,  
        preprocessor=None, stop_words=None, max_features=5000)  
  
479: print "Training Random forest..."  
  
484: print "Cleaning and parsing \n"  
  
486: clean_test_reviews.append(" ".join(KaggleWord2VecUtility.review_t  
        o_wordlist(test['Snippet'][i], True)))  
  
490: print "Predicting test labels...\n"  
  
492: output = pd.DataFrame(data={"Accuracy": "", "Sentiment":result,  
        "Ticker":test["Ticker"], "Date":test["Date"]})  
  
494: print "Wrote results to randomForestResults.csv"  
  
561: temp_date = datetime.datetime.strptime(endDate, "%Y-%m-%d")  
  
passed: False
```

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

find input{format/final}

---

passed: False

floats

---

123: \begin{figure}[htb]

209: \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}

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

244: \begin{figure}[htb]

332: \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}

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

```
441: \begin{figure}[htb]
497: \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}
500: Figure \ref{c:sentiment} shows the entire code to train on the
      dataset provided by the historical stock data and Google Alert
      sentiments.
504: \begin{figure}[htb]
593: \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}
596: Figure \ref{c:result} shows the code to combine the data produced
      by the random forest analysis and combine it with available
      historic stock price data.
```

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

```
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: algorithm and above may be used improperly

426: 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

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 Humblet, 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.prb.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] Malaria 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://medicaleconomics.modernmedicine.com/medical-economics/news/here-s-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.22012>
- [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://ppq.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/optum3/en/images/infographics/Game\\_changer\\_Track\\_Two\\_04\\_Data\\_Rich\\_Insight\\_Poor\\_Infog/Images\\_2016.pdf](https://cdn-aem.optum.com/content/dam/optum3/optum3/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.2014.02.015>

- [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](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>
- [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
```

```
[2017-12-10 13.51.49] 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 10028 report.tex
wc 327 project 11 10779 report.pdf
wc 327 project 11 2080 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
```

```
passed: True
```

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

```
4: \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 Darrow, David Reich, Andrew Kasarskis, Nicholas P Tatonetti, 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.]

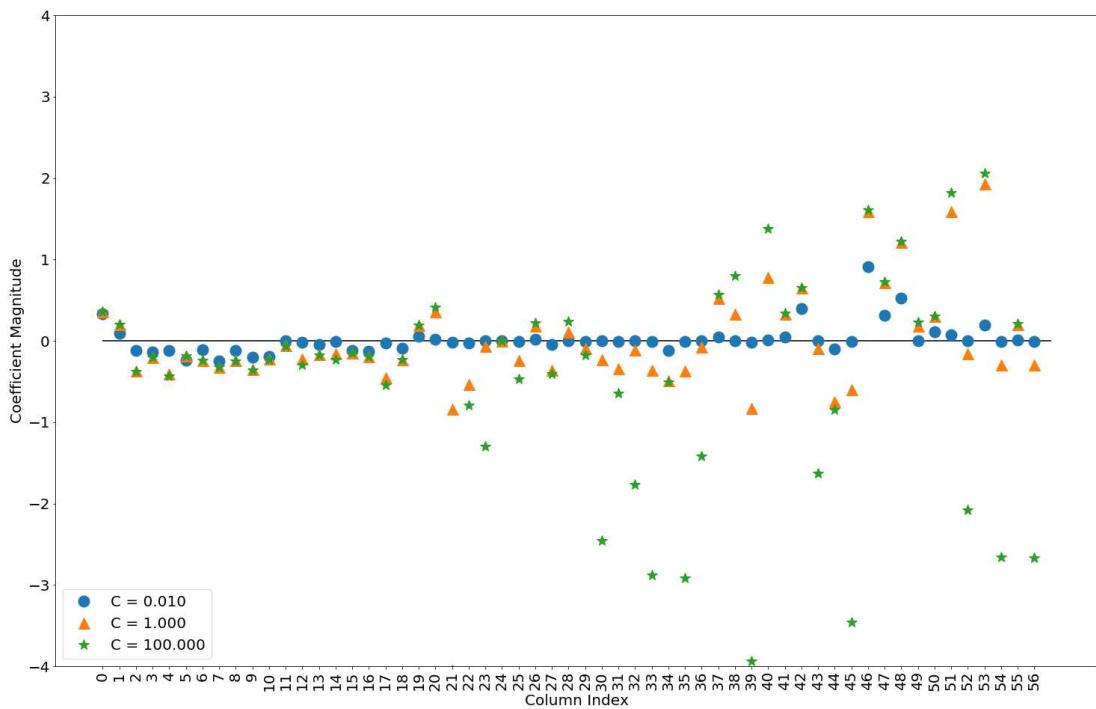
[Figure 2 about here.]

[Figure 3 about here.]

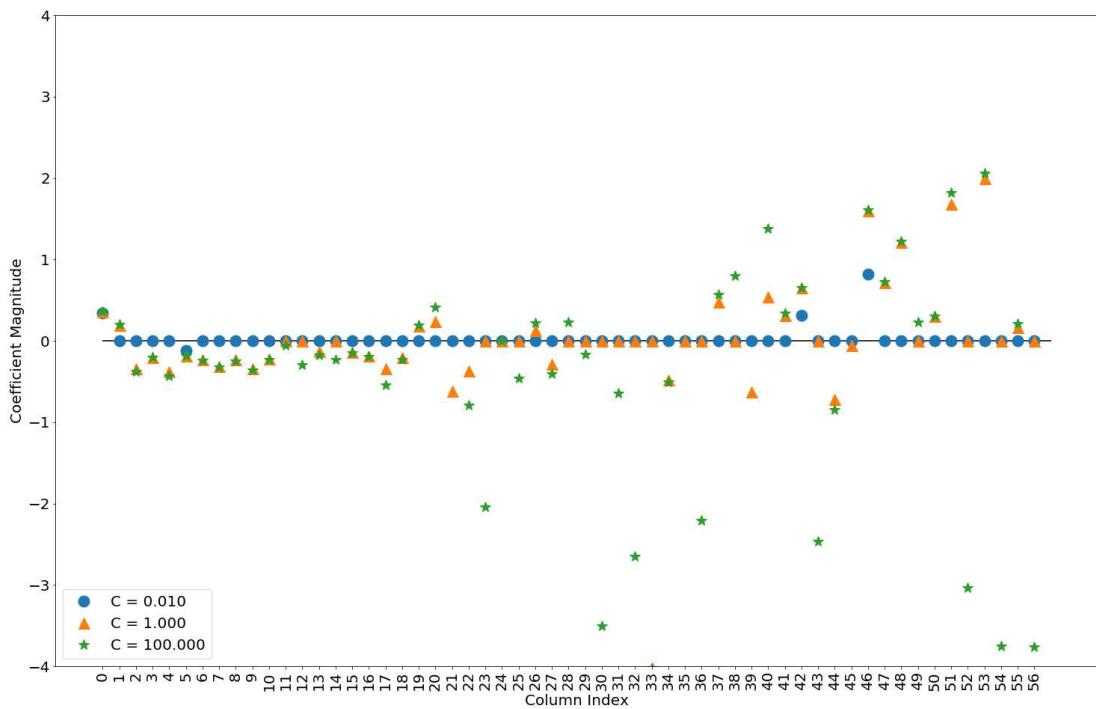
[Figure 4 about here.]

LIST OF FIGURES

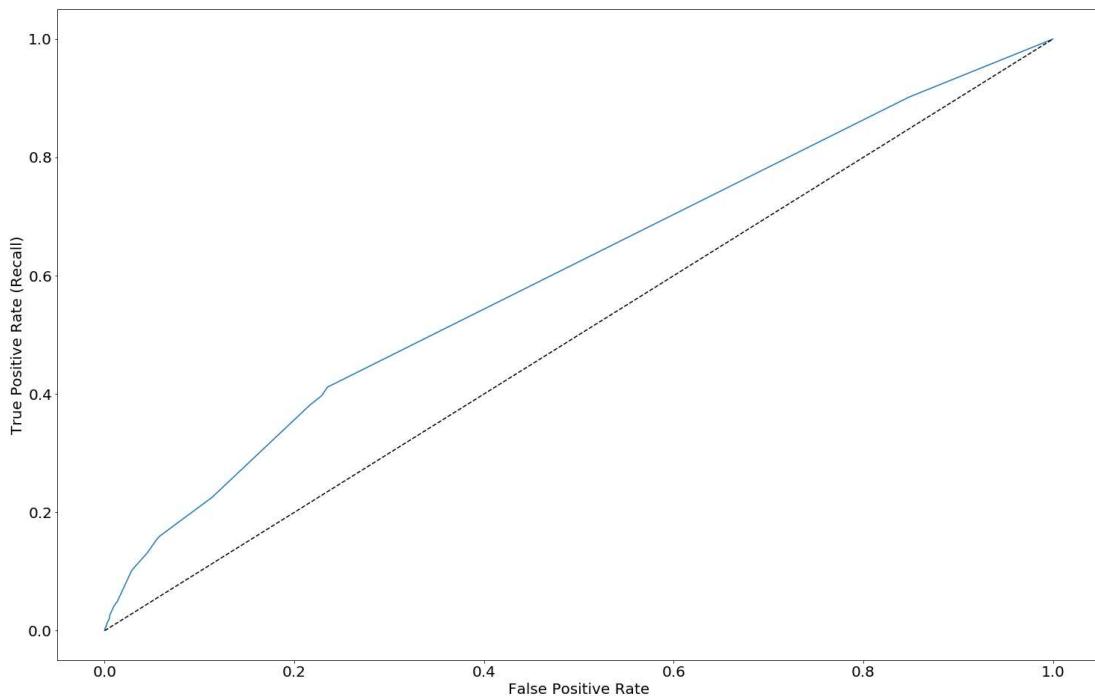
1	Logistic Regression Weights By C-Value, L2 Regularization	9
2	Logistic Regression Weights By C-Value, L1 Regularization	10
3	ROC Curve, L1 Regularization, C = 0.01	11
4	ROC Curve, L1 Regularization, C = 100.0	12



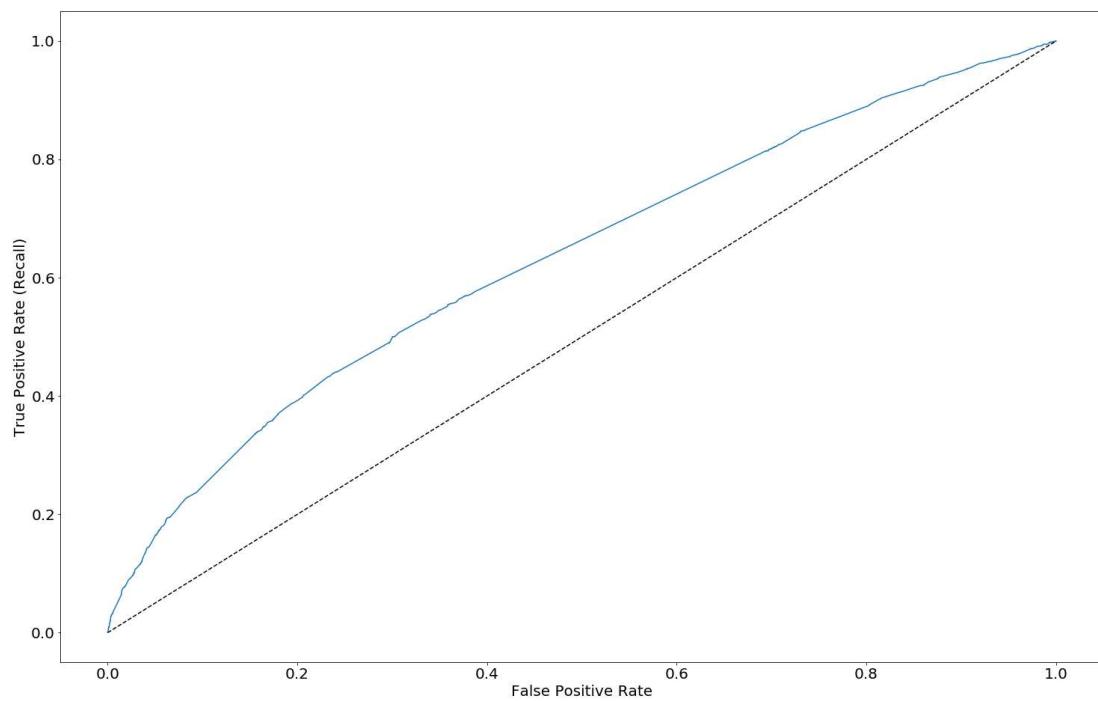
**Figure 1: Logistic Regression Weights By C-Value, L2 Regularization**



**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
```

---

```
[2017-12-10 13.52.19] 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 6934 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}  
-----  
  
4: \input{format/i523}  
  
passed: True  
  
find input{format/final}  
-----  
  
passed: False  
  
floats  
-----  
  
196: 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.

210: 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.

236: \begin{figure}[!ht]  
237: \centering\includegraphics[width=\columnwidth]{images/weightsl2.png}  
238: \caption{Logistic Regression Weights By C-Value, L2 Regularization}\label{f:weightsl2}  
241: \begin{figure}[!ht]  
242: \centering\includegraphics[width=\columnwidth]{images/weightsl1.png}  
243: \caption{Logistic Regression Weights By C-Value, L1 Regularization}\label{f:weightsl1}  
246: \begin{figure}[!ht]  
247: \centering\includegraphics[width=\columnwidth]{images/roccurve001.png}  
248: \caption{ROC Curve, L1 Regularization, C = 0.01}\label{f:roccurve001}  
251: \begin{figure}[!ht]  
252: \centering\includegraphics[width=\columnwidth]{images/roccurve100.png}  
253: \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 is 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 Lives. 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] Liaison. 2017. Value Based Healthcare - The patient is the Center but Data is the Key. Web page as blog. (June 2017). <https://www.liaison.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/20-must-see-stats>
- [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 Act 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
```

```
=====
report.bib:104: @Misc{www-google-liason,
report.bib:113: @Misc{www-google-datameer,
report.bib:122: @Misc{www-google-pred,
report.bib:131: @Misc{www-google-EHR,
report.bib:140: @Misc{www-google-rock,
report.bib:149: @Misc{www-google-elec,
report.bib:158: @Article{springer,
report.bib:191: @Misc{www-google-wikitele,
report.bib:200: @Misc{www-google-forbes042015,
report.bib:209: @Misc{www-google-datapine,
report.bib:20: @Misc{www-google-Christian,
report.bib:218: @Misc{www-google-ASC,
report.bib:227: @Misc{www-google-wikievi,
report.bib:236: @Misc{www-google-trillion,
report.bib:245: @Misc{www-google-geno,
report.bib:254: @Misc{www-google-transparent,
report.bib:263: @Misc{www-google-ACA,
report.bib:272: @Misc{www-google-cost,
report.bib:281: @Misc{www-google-pacient,
report.bib:290: @Misc{www-google-drug,
report.bib:299: @Misc{www-google-hadoop,
report.bib:29: @Misc{www-google-McDonald,
report.bib:308: @Misc{www-google-reas,
report.bib:317: @Misc{www-google-wall,
report.bib:326: @Misc{www-google-chicago,
report.bib:335: @Misc{www-google-sandiego,
report.bib:344: @Misc{www-google-admin,
report.bib:353: @Misc{www-google-data,
report.bib:362: @Misc{www-google-pharmrd,
report.bib:38: @Misc{www-google-CIO,
```

```
report.bib:47: @Misc{www-google-HlthCat,
report.bib:56: @Misc{www-google-PBS0,
report.bib:65: @Misc{www-google-digit,
report.bib:74: @Article{milbank,
report.bib:86: @Misc{www-google-pracfus,
report.bib:95: @Misc{www-google-AC0,
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-12-10 13.52.25] 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 1.0s.
./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 8954 report.tex
wc 332 project 10 9324 report.pdf
wc 332 project 10 1543 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

```
4: \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.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