Use Cases in Big Data Software and Analytics

Vol. 1, Fall 2017

Bloomington, Indiana

Monday 4th December, 2017, 22:03

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

Contents

1 Preface **9**

1.0.1 Disclaimer 9

1.0.2 Citation 9

1.1 List of Papers 10

2 Biology **13**

3 Business **13**

2 hid106 Status: 0%

A Music Recommendation System

Shiqi Shen, Qiaoyi Liu 13

3 hid109 Status: 100%

A Music Recommendation System

Shiqi Shen, Qiaoyi Liu 13

4 hid302 Status: 100%

Recipe Ingredients Analysis

Sushant Athaley 19

5 hid310 Status: 100%

Gerrymandering Detection Using Data Analysis

Kevin Duffy 62

4 Edge Computing **62**

5 Education **62**

6 Energy **62**

7 Environment **62**

8 Government **62**

9 Health **62**

6 hid327 Status: 99%

How Big Data will Help Improve People's Health Worldwide

Paul Marks 62

7 hid335 Status: 100%
Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics
Sean Shiverick 78

10 Lifestyle **110**

11 Machine Learning **110**
8 hid209 Status: Dec 04 17 100%
Comparison between different classification algorithms in Digit Recognizer
Han, Wenxuan, Liu, Yuchen, Lu, Junjie 110
9 hid343 Status: 100 %
Income Prediction Using Machine Learning Techniques
Borga Edionse Usifo 135

12 Media **162**
10 hid208 Status: Dec 04 17 100%
Big Data Analytics on Influencers in Social Networks
Jyothi Pranavi Devineni 162

13 Physics **162**

14 Security **162**
11 hid316 Status: 100%
Edge Analytics for Weather Monitoring and Forecasting
Robert Gasiewicz 162

15 Sports **162**

16 Technology **162**
12 hid306 Status: 100%; 12/3/2017
Predicting Housing Prices - Kaggle Competition
Murali Cheruvu, Anand Sriramulu 162
13 hid315 Status: 80% - finishing up .txt
TBI - A Data Driven Journey Beyond Contact Sports... Putting Data In The Drivers Seat
Garner, Jeffry 187
14 hid337 Status: Dec 04 17 100%
IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)
Ashok Reddy Singam, Anil Ravi 187

17 Text **202**

18 Theory **202**

19 Transportation **202**

20 TBD **202**

15 hid101 Status: Dec 04 17 0%
TBD
Huiyi Chen 202
16 hid102 Status: Dec 04 17 0%
None
Gregor von Laszewski 202
17 hid104 Status: 0%
TBD
Jones, Gabriel and Millard, Mathew 202
18 hid105 Status: Dec 04 17 0%
None
Gregor von Laszewski 202
19 hid107 Status: Dec 04 17 0%
None
Gregor von Laszewski 202
20 hid201 Status: 100%
IoT Application Using MQTT and Raspberry Pi Robot Car
Arnav, Arnav 202
21 hid202 Status: 0%
NA
Himani Bhatt 222
22 hid203 Status: 100% Dec 4, 2017
Big Data Analytics on Food Products Around the World
Chandwani, Nisha and Vegi, Karthik 222
23 hid204 Status: 0%
Big Data in Safe Driver Prediction
Wang, Jiaan and Chaturvedi, Dhawal 222
24 hid205 Status: 0%
Benchmarking a BigData Docker deployment
Gregor von Laszewski 222
25 hid210 Status: Dec 04 17 0%
None
Gregor von Laszewski 222
26 hid211 Status: unkown
Continuous motion tracking using Deep Neural Networks and Recurrent Neural Networks
Khamkar, Ajinkya 222
27 hid212 Status: Dec 04 17 0%
None
Gregor von Laszewski 222
28 hid214 Status: 0%
Benchmarking a BigData Docker deployment
Junjie Lu 222
29 hid216 Status: Dec 04 17 0%
None
Gregor von Laszewski 222

30 hid218 Status: unkown
None
Geng Niu 222
31 hid219 Status: Dec 04 17 0%
Fake Review Detection on Yelp Dataset
Gregor von Laszewski 222
32 hid224 Status: 0%
Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks
Rawat, Neha 230
33 hid225 Status: Dec 04 17 0%
None
Gregor von Laszewski 230
34 hid230 Status: unkown
big data with natural language processing
Yuanming Huang 230
35 hid231 Status: 99%
Big Data Analytics on Food Products Around the World
Vegi, Karthik and Chandwani, Nisha 230
36 hid232 Status: 0%
Big Data and Hearing Disabilities
Rahul Velayutham 254
37 hid234 Status: 70%
Big Data Applications in the Travel Industry and its Potential in Improving Travel Accessibility
Weixuan Wang 263
38 hid235 Status: unkown
Big Data analytics in predict house price
Yujie Wu 269
39 hid237 Status: 10%, Dec 4, 2017
Analyzing everyday challenges of people with visual impairments
Tousif Ahmed 269
40 hid304 Status: Dec 04 17 100%
How Far have Space Walks Walked
Ricky Alan Carmickle 273
41 hid308 Status: 0%
Benchmarking a BigData Docker deployment
Pravin Deshmukh 273
42 hid312 Status: not yet started
To be decided
Neil Eliason 273
43 hid313 Status: 90%
The Impact of Clinical Trial Results on Pharmaceutical Stock Performance
Tiffany Fabianac 273
44 hid314 Status: Dec 04 17 0%
None
Gregor von Laszewski 283

45 hid318 Status: Dec 04 17 0%
None
Gregor von Laszewski 283
46 hid319 Status: 30%
Facial Recognition and Object Detection using Raspberry Pi Robot Car
Mani Kumar Kagita 283
47 hid320 Status: 100% Dec 03 2017
Real Estate Big Data Analysis
Elena Kirzhner 283
48 hid321 Status: Dec 04 17 0%
None
Gregor von Laszewski 283
49 hid323 Status: Dec 04 17 0%
None
Gregor von Laszewski 283
50 hid324 Status: Dec 04 17 95%
Big Data Analytics in factors affecting Bitcoin
Ashok Kuppuraj 283
51 hid326 Status: unkown
None
Mohan Mahendrakar 283
52 hid329 Status: 100% Dec 4
Big Data and The Customer Experience Journey
Ashley Miller 283
53 hid331 Status: Dec 04 17 99%
Big Data Applications in Predicting Hospital Readmissions
Tyler Peterson 305
54 hid332 Status: 100%
Big Data Analytics to Reduce Health Care in the United States
Judy Phillips 322
55 hid333 Status: 100%
IoT and Big Data Analytics for Equipment Predictive Health Management type: latex
Anil Ravi, Ashok Reddy Singam 341
56 hid334 Status: Dec 04 17 100%
The Intersection of Big Data and IoT
Peter Russell 341
57 hid336 Status: Dec 04 17 0%
None
Gregor von Laszewski 353
58 hid338 Status: Dec 04 17 0%
None
Gregor von Laszewski 353
59 hid339 Status: Dec 2 2017 100%
Diagnosis of Coronary Artery Disease Using Big Data Analysis
Hady Sylla 353

60 hid340 Status: Dec 04 17 0%
None
Gregor von Laszewski 353
61 hid341 Status: 0%
Not submitted
Tibenkana, Jacob 353
62 hid342 Status: 0%
TBD
Nsikan Udoyen 353
63 hid346 Status: unkown
NOt submitted
Zachary Meier 353
64 hid348 Status: 80%
Not submitted
Budhaditya Roy 354

Chapter 1

Preface

1.0.1 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.

1.0.2 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},
}
```

1.1 List of Papers

HID	Author	Title
101	Huiyi Chen	TBD
0	Gregor von Laszewski	None
104, 216	Jones, Gabriel and Millard, Mathew	TBD
0	Gregor von Laszewski	None
109, 106	Shiqi Shen, Qiaoyi Liu	A Music Recommendation System
0	Gregor von Laszewski	None
109, 106	Shiqi Shen, Qiaoyi Liu	A Music Recommendation System
hid111	error: yaml	A Music Recommendation System
hid201	error: yaml	IoT Application Using MQTT and Raspberry Pi Robot Car
202	Himani Bhatt	NA
203 and 231	Chandwani, Nisha and Vegi, Karthik	Big Data Analytics on Food Products Around the World
233 and 204	Wang, Jiaan and Chaturvedi, Dhawal	Big Data in Safe Driver Prediction
205	Gregor von Laszewski	Benchmarking a BigData Docker deployment
208	Jyothi Pranavi Devineni	Big Data Analytics on Influencers in Social Networks
209, 213, 214	Han, Wenxuan, Liu, Yuchen, Lu, Junjie	Comparison between different classification algorithms in Digit Recognizer
0	Gregor von Laszewski	None
211	Khamkar, Ajinkya	Continuous motion tracking using Deep Neural Networks and Recurrent Neural Networks
0	Gregor von Laszewski	None
hid213	error: yaml	None
214	Junjie Lu	Benchmarking a BigData Docker deployment

215	Mallala, Bharat	Big Data Analytics on Influencers in Social Networks
0	Gregor von Laszewski	None
218	Geng Niu	None
219	Gregor von Laszewski	Fake Review Detection on Yelp Dataset
224	Rawat, Neha	Big Data Analytics in Detection of DDoS (Distributed Denial-of-Service) attacks
0	Gregor von Laszewski	None
hid228	error: yaml	None
hid229	error: yaml	None
230	Yuanming Huang	big data with natural language processing
231	Vegi, Karthik and Chandwani, Nisha	Big Data Analytics on Food Products Around the World
232	Rahul Velayutham	Big Data and Hearing Disabilities
233 and 204	Wang, Jiaan and Chaturvedi, Dhawal	Big Data in Safe Driver Prediction
hid234	error: yaml	Big Data Applications in the Travel Industry and its Potential in Improving Travel Accessibility
235	Yujie Wu	Big Data analytics in predict house price
hid236	error: yaml	Big Data analytics in predict house price
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	Benchmarking a BigData Docker deployment
hid309	error: yaml	Benchmarking a BigData Docker deployment
310	Kevin Duffy	Gerrymandering Detection Using Data Analysis
hid311	error: yaml	Gerrymandering Detection Using Data Analysis
None	Neil Eliason	To be decided
313	Tiffany Fabianac	The Impact of Clinical Trial Results on Pharmaceutical Stock Performance
0	Gregor von Laszewski	None
315	Garner, Jeffry	TBI - A Data Driven Journey Beyond Contact Sports... Putting Data In The Drivers Seat
316	Robert Gasiewicz	Edge Analytics for Weather Monitoring and Forecasting
0	Gregor von Laszewski	None
319	Mani Kumar Kagita	Facial Recognition and Object Detection using Raspberry Pi Robot Car
320	Elena Kirzhner	Real Estate Big Data Analysis
0	Gregor von Laszewski	None
0	Gregor von Laszewski	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?
218	Mohan Mahendrakar	None
327	Paul Marks	How Big Data will Help Improve People's Health Worldwide
hid328	error: yaml	How Big Data will Help Improve People's Health Worldwide

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
333, 337	Anil Ravi, Ashok Reddy Singam	IoT and Big Data Analytics for Equipment Predictive Health Management type: latex
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
0	Gregor von Laszewski	None
337, 333	Ashok Reddy Singam, Anil Ravi	IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)
0	Gregor von Laszewski	None
332	Hady Sylla	Diagnosis of Coronary Artery Disease Using Big Data Analysis
0	Gregor von Laszewski	None
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: Then, Now, and Beyond
346	Zachary Meier	NOt submitted
347	Jeramy Townsley	Mapping Police Killing of Citizens in the United States
348	Budhaditya Roy	Not submitted

A Music Recommendation System

Shiqi Shen
Indiana University Bloomington
1575 S Ira St
Bloomington, Indiana 47401
shiqshen@indiana.edu

Qiaoyi Liu
Indiana University of Bloomington
3209 E 10th St
Bloomington, Indiana 47408
ql30@umail.iu.edu

ABSTRACT

Recently, people tend to use auto play-next while using a cloud-based music app. In this case, content recommendation is at the heart of most subscription-based media stream platforms. A good recommendation system based on a huge number of historical records and metadata can vastly enhance user experience and increase user engagement. In our project, we would build machine learning models to predict what songs are going to be listened next week. The training data would be a copy of user listening history during a month period and predict what songs a set of predetermined users would listen to during the next week.

KEYWORDS

i423, hid109, Big Data; Amazon; Customers; Pricing; Dynamic, Internet, application

1 INTRODUCTION

The ability to generate and exchange information has increased tremendously over the recent past. This growth is driven by the easy availability and affordability of the computing as well as the ubiquity of the internet [3]. In the current businesses world, almost everything is conducted electronically. There is a lot of information exchange and engagement over the internet, as well as selling and buying of products. Amazon is one of the leading giants in the application of big data. The firm is one of the pioneers of e-commerce, and one of its most outstanding innovations in this domain is the personalized recommendation system. The foundation of the system is big data, which is usually collected from the customers. The firm has received various coveted awards due to its excellent innovations and application of big data [3]. The firm has leveraged big data in the recent past to enhance its performance as well as service delivery to the customers. Together with other major firms in the internet services industry, Amazon acknowledged the significance of big data in the initial years of 2000, and then immediately focused on adequately using the big database of clients shopping on its online platforms.

Big data operates on the concept of the power of suggestion, as fronted by psychologists. They claim that by putting something that an individual may like in front of them, then they may have strong desire to purchase it. Amazon employed this philosophy by leveraging their customer data and transforming its system into a high powered one that is focused on the customer. The firmfis systems have been getting better by the day and expected to be even more superior in the near future.

2 PRODUCT RECOMMENDER SYSTEM

In the recent past, Amazon has moved from operating as a pure e-commerce firm to a major player in the internet services industry, with focus on offering a wide variety of services to both individuals as well as companies. The firm started to shift its focus on big data and started the journey to transition from a typical online retailer into one a major force in the realm of big data. Around 2000, the company, along with other internet firms such as Google, Yahoo, and Twitter realized that they had voluminous data about their customers, which could be put used to improve their performance. Although the other firms did not initially concentrate majorly on big data, Amazon swiftly moved to take advantage of the invaluable database of individuals who used its e-commerce platforms around the world to shop. The team charged with the responsibility of recommending the products to the customers came up with innovative strategies that the firm could make use of the data collected by the firm about their customers. The end result of the move was a huge success in big data, which revolutionized how the company did business.

As a major player in the e-commerce domain, the success of Amazon was always pegged on availing the right products to the customers. The efficacy of providing the right products for the customers in turn largely depended on a proper understanding of the needs of the consumers. A proper market research was necessary in order to understand the customerfis needs and tastes. Since it was founded, Amazon has created a name for itself because of its superior product recommender system, which suggests products to consumers on the basis of their last purchase. The major driving force behind the recommender system is the data gathered from the customers.

The product recommender system is essential for the personalization of each customerfis experience when they are shopping in the firmfis online store [6]. The firm employs collaborative filtering and clustering algorithms to classify clients on the basis of preferences. Customers are grouped on the basis of same search as well as collaborative filtering between items. Content-based search employs the shopping history of customers and item ratings to establish a search query capable of finding other items that match the tastes of consumers. For instance, if a customer purchases a book, the product recommender systems will suggest books from the same author, publisher, or subject area. The product recommendations are not only used by the company in the online stores, but it also doubles up as a marketing tool useful in conducting email campaigns. There is a recommendation link that enables shoppers to filter products by several criteria depending on the items that they have in their shopping carts.

3 BIG DATA FOR DYNAMIC PRICING

Dynamic pricing entails the use of big data such as clickstreams, purchase history, cookies, etc. to offer customized discounts to customers or to alter the prices of items being sold dynamically. The technology enables the real-time price customization for an item to suit a specific customer. This explains why it is sometimes possible for two different sets of customers to buy the same item at different prices from the same online store [5]. Despite the immense benefits of this technology, some customers may always feel discriminated against due to the price differences. Amazon has successfully used the power of big data to implement a price discrimination system. For example, there was an incident in which some Amazon customers were aggravated about price variations of a certain DVD. One of the customers noted that there was a difference of nearly two points five dollars in the price if the cookers were deleted from the computer. Price discrimination was also experienced in the sale of a product known as Diamond Rio MP3 Player.

Big data also enables price optimization. This enables the firm to manage the prices of commodities and grow its profits by twenty-five percent annually. Several factors are used to set the prices of commodities. Some of them are: activity of the customer on the firmfis shopping portal, availability of the product, competitorfis prices, order history, item preferences, and the anticipated profit margin [5]. The prices are normally refreshed every ten minutes as big data become updated. Due to this, Amazon provides customers with discounts on best-selling commodities and accrue large profit margins on the items that are less popular with customers.

4 BIG DATA AND CUSTOMER SERVICE

Big data is also extensively being used for customer service at Amazon. The acquisition of Zappos has often been viewed as a major element in the same. Since it was founded, Zappos has enjoyed a good reputation for the excellence in customer service and was usually viewed as a world leader in this domain. Much of the success can be attributed to their advanced relationship management systems which extensively employed their own customer data. After the acquisition of the form in 2009, the procedures were integrated together with those of Amazon. Todayfis business environment is changing at a rapid rate, and consumers are also using their voices faster. Within a few moments after undergoing a bad experience, customers can swiftly move into social media and spread the news about their negative experience [4]. The only strategy for an organization to survive under such conditions is to employ the power of analytic to streamline and shorten the response time, as well as fix the customer support issues. The customers of the present day are not only looking for a product that works, but also one that is personalized and able to recognize their interests and save them time.

5 ONE CLICK ORDERING

Amazon used big data to create one-click ordering. This feature is activated automatically when the customer places his first order, enters a shipping address as well as a method of payment. When using the one-click feature, the customer is given thirty minutes to change his mind about the particular purchase. This system was created on the premise that a simplified path to purchase would increase conversion rates. Since the introduction of the technology, the firmfis revenues have increased year after year. The significance of this application pushed the company to patent it to prevent other companies from using it without authorization. Reorganizing the purchase process is currently one of the most significant differentiates in the current marketplace. The service enables users to make payments without having to exchange cards or money physically. Amazon has also greatly benefited from impulse buying, which is accelerated by one-click buying. Research has shown that the largest percentage of people normally purchase things they donfit require or did not plan to purchase in the first place [2].

6 USING BIG DATA TO SUPPORT OTHER COMPANIES

Amazon also uses its big data platform to support and help other companies improve their operations. Organizations can employ AWS toolkit provided by Amazon to create scalable big data applications that have the capacity to improve business performance [6]. Besides, they would be able to secure these applications easily without the need to spend on expensive infrastructure and hardware. The big data applications including data warehousing, clickstream analytic, fraud detection, internet of things, and several others are delivered via cloud computing. Hence, there is no need for an organization to incur additional costs in setting up a data center. The Amazon web services can enable companies to analyze spending habits, customer demographics, and other related information to enable them effectively cross-sell some of the firmfis products in patterns similar to Amazon. That is to say that the retailers will also be able to stalk their customers, recommend products to them, and improve their customer experience.

7 BIG DATA TECHNOLOGIES

Amazon EMR: This technology offers a managed Hadoop framework that simplifies and hastens the processing of huge amounts of data across scalable Amazon EC2 instances. Amazon EMR also supports other common distributed frameworks including HBase, Apache Spark, Flink, and Presto [1]. Besides, it reliably and safely handles a wide range of big data use cases, such as web indexing, log analysis, financial analysis, machine learning, and bioinformatics.

Amazon Athena: It denotes an interactive query service that simplifies data analysis in Amazon S3 via standard SQL. Since it is service less, one only pays for the queries they run and there is no infrastructure to be managed [1]. The technology is quite straightforward and delivers results within the shortest time possible. Moreover, it does not require complex ETL jobs to prepare data for analysis.

Amazon Kinesis Firehouse: This is one of the simplest methods to import streaming data into Amazon Web Services. The technology can be used to gather, transform, and import streaming data into Amazon S3, Amazon Kinesis analytic, and Amazon Redshift, to permit instant analytic with the current BI tools and dashboards currently being used. It is a comprehensively managed service that can expand automatically with the increase in data throughput.

8 CONCLUSION

Big data has grown tremendously in the recent past. The growth has been accelerated majorly by the increased accessibility of computing devices as well as the ubiquity of the internet. Being one of the pioneers of e-commerce, Amazon has extensively employed big data to improve its performance. Big data has been used to create recommender systems, implement dynamic pricing, streamline and improve the customer experience, and support other companies. The system recommends products to customers based on their purchase history and enables them to filter the products list based on certain criteria. The company continues to enhance its big data applications with a view to creating a loyal customer base.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support to write this paper as well as TAs' helpful suggestions on this paper.

REFERENCES

[1] Amazon. 2017. Big Data on AWS. (2017). https://aws.amzon.com/big-data/
[2] Roy F Baumeister. 2002. Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of consumer Research* 52, 4 (2002), 670–676.
[3] Marc L Berger & Vitalii Doban. 2014. Big data, advanced analytics and the future of comparative effectiveness research. *Journal of company effectiveness research* 3, 2 (2014), 167–176.
[4] Randal E. Bryant & Randy H. Katz & Edward D. Lazowska. 2008. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. (2008). https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf
[5] Benjamin Reed Shiller. 2014. First-Degree Price Discrimination Using Big Data. (2014). http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data_Jan_18,_2014.pdf
[6] Hsinchun Chen & Roger H L Chiang & Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012), 1165–1188.

```
bibtext report
===========================================

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

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


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


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


latex report
===========================================

        [2017-12-04 22.00.16] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
```

```
Typesetting of "report.tex" completed in 0.9s.

================================================================================
Compliance Report
================================================================================


name: Shiqi Shen
hid:  109
paper1: complete 100% Oct 27th
paper2: complete 100% Nov 4th
project: 100%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

3
wc 109 project 3 2065 report.tex
wc 109 project 3 2107 report.pdf
wc 109 project 3 199 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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------


bibtex
--------------------------------------------------------------------------------


label errors


bibtex errors

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

```
Database file #1: report.bib

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------

non ascii found 8217
non ascii found 8217
non ascii found 8217
non ascii found 8217
non ascii found 8217
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
```

Recipe Ingredient Analysis

Sushant Athaley
Indiana University
sathaley@iu.edu

ABSTRACT

Food is the unavoidable part of day to day of human life. Ingredients play a major role or are the basic requirement in preparation of any kind of food. We can find the humongous list of ingredients getting used across globally along with other details which constitute to big data. We explore ingredients getting used in various recipes across the globe to understand most used ingredient, key ingredients of various cuisine and the relationship between the ingredients to find out closely related ingredients which can always provide great dish if used together.

KEYWORDS

i523, hid302, big data, ingredient, recipe, analysis, python, gephi

1 INTRODUCTION

Ingredients are vital for human existence as well as for food or restaurant industry. We use it every day for cooking and food industry uses it to produce consumable for their customers. Ingredient inspires chefs to come up with new culinary artistry. So what do we know about this essential element of the life and what data tell us? Ingredients come in different size, color, shape, flavor, nutrition, taste, texture, grows in specific weather conditions and this provides a great opportunity for various analysis which can be useful for the human being as well as business industries. So main focus of this study is on the ingredients used in various recipes across the cuisines. This study evaluates recipe ingredient dataset from Kaggle [7] to analyze most used ingredients, key cuisine ingredients and ingredient relationship.

This study is organized as follows, section *Ingredient* defines ingredient and it's various characteristics. section *Ingredeint Analytics* describes various analytics which can be performed on the ingredient with some examples. Section *Project* describes the aim of this study. Section *technologies* provides information on the tools and technologies used for this project. Section *Methodology* covers overall process carried out in this project. Section *Dataset* describes data structure used along with loading process and data findings. Section *Analysis and Findings* describes various analysis carried out on the data and the visual representation of the analysis. Section *Shortcomings* captures shortcomings of the project. Section *Future Work* talks about what else can be done with this dataset which is not covered in the current scope of the project. Section *Conclusion* concludes the study.

2 INGREDIENT

Food is defined as "Edible or potable substance (usually of animal or plant origin), consisting of nourishing and nutritive components such as carbohydrates, fats, proteins, essential mineral and vitamins, which (when ingested and assimilated through digestion) sustains life, generates energy, and provides growth, maintenance, and health of the body" [2]. Thus food is the basic necessity for human for the sustainability. Food can be eaten raw, cooked or processed. As human race evolved over the period of time, the way we eat food is also evolved. Food cooking is just not the basic necessity but its an art and science in today's era. Food preparation consists of various cooking techniques, tools, and ingredients to make it palatable or edible by humans. The ingredient is by far the most important part of any food or recipe preparation. The recipe consists of the list of ingredients and the set of instruction to cook particular food dish [5]. An ingredient is defined as "Any of the foods or substances that are combined to make a particular dish" [9]. Ingredients impart various flavors, aroma, texture, and color to the cooking dish. Ingredients are mostly derived from vegetables, fruits, nuts, grains, living organisms, herbs, flowers, and spices. It comes in both solid and liquid forms. Another characteristic of ingredients is the nutritional value they provide which is essential for the human body.

3 INGREDIENT ANALYTICS

Ingredients characteristics and the combination of other related data provides various opportunities to analyze ingredient in different ways. Analysis of the flavors present in ingredient can provide us with the categorization of the different ingredient by the flavor profile which can be helpful in deciding substitute ingredient if a certain ingredient is not present or pairing ingredient from different flavor categories to construct the dish as per the taste required [1]. This analysis also helps to understand which ingredients cannot be used together. A similar analysis is carried out to correlate ingredient across recipes to come up with top 50 combinations of ingredients which can be used together [6]. Flavourspace application provides functionality to search recipe based on the ingredients, suggests alternate ingredient if not present, adjust the recipe as per the taste which is a good example of big data analytics in food industry [10]. Foodpairing application takes another approach to form the connection between unfamiliar ingredients and provides information on how to use such ingredient to make a dish, this is very helpful in terms of sustainability as we can use ingredient which is ample available but not in use due to the absence of information on using such ingredients [8].

Another study conducted on most used ingredient provides insight that sugar, oil, pepper, and salt are most commonly occurring ingredient, among spices clove, in vegetable onion, garlic , and tomatoes, butter in milk product, eggs followed by chicken in the animal product are the most used ingredient in the categories [3]. This information can help in better planning and sourcing of such ingredients which are in high demand.

Ingredient nutrition analysis can help find out nutrition of the food prepared by those ingredients. This would be helpful in menu planning where nutrition information is the key factor such as school, hospitals or any other dietary program [4].

Recipe cost is calculated by including the cost of the ingredient used in that recipe. Ingredient cost as per the quantity used in recipe provides base information to calculate the price of any recipe. This ingredient cost analysis provides an avenue to reduce the cost of the recipe by using substitute ingredient of lesser cost. This can also help in household budget to keep in check as well as make restaurant industry profitable.

Ingredient used in recipe can provide insight into type of weather received by that cuisine as ingredient can grow in certain weather condition. This can help chef locally source the ingredient and maintain local agriculture sustainability.

4 PROJECT

This project study is conducted to analyze ingredients getting used in various recipes across the cuisines to find out

- Most used ingredients across cuisines or globally
- Key ingredients used by cuisines
- Ingredient relationship or connection to understand the related ingredients

4.1 Technologies

Technologies and tools used in this projects are

- Python version 3.6 is used for data load and processing
- Gephi 0.9.2 for visualization
- Spyder 3.0 as a Python IDE

4.2 Methodology

The first step was to source the data. We were interested in the dataset which provides recipe information along with the ingredient used in the recipe. Since we wanted to analyze distribution across cuisines, data should also contain cuisine tagging. This dataset can be generated by pulling recipe data from various online applications or pick from publicly available datasets. We finalized publicly available dataset at Kaggle application satisfying need for this project.

Figure 1 shows methodology used for this project to analyze ingredient data.

[Figure 1 about here.]

DataSet is loaded through Python script and further processed to clean the data. This cleaned data then processed to analyze ingredient distribution across cuisine and per cuisine. Gephi software is used to analyze the relationship and to find out the ingredient modularity. The Python script is used to create the network files required by the Gephi tool. Gephi requires Nodes and relationship in terms of Edges between the nodes for the analysis. The Python script is used to create Node and Edges file in excel format so that it can be imported into Gephi. Distinct ingredients used in recipe becomes the nodes. Edges or relationship between ingredients is derived by relating ingredients appearing in the same recipe. All ingredient in the same recipe is considered related to each other. Network files created by Python are imported in Gephi to produce the graph for the visualization. Gephi tools data laboratory is used to clean up the data and filters are applied to provide usable network visualization.

4.3 Dataset

The dataset for this study is sourced from Kaggle application [7]. This dataset is publicly available and featured in "What's Cooking?" competition. This dataset is in JSON format and of 12MB size. This dataset contains recipe id, cuisine and list of ingredients as described in Figure 2.

[Figure 2 about here.]

This dataset contains total 39774 recipes across various cuisines. We used two different methods to load this data. Cuisine and ingredient analysis is done by loading data into *pandas dataframe* and to analyze ingredient relationship data has been loaded into *json* object. Figure 3 shows the code for data loading used in this project.

[Figure 3 about here.]

Ingredient extraction from the data structure and processing was challenging as ingredients are listed comma separated for each recipe. Also, ingredient list can vary by recipe and there is no proper structure. Another issue with the ingredient list is ingredient appears in various forms but it's the same ingredient which gives duplicate data. For example, salt appears as salt, kosher salt, Morton Salt, sea salt, table salt, Himalayan salt, fine sea salt, low sodium salt, fine salt. This is the same ingredient but come across in recipe as a different ingredient and getting counted as a separate ingredient in the analysis. Some ingredients are listed along with measures like (10 oz.) frozen chopped spinach, (10 oz.) frozen chopped spinach, thawed and squeezed dry, (14.5 oz.) diced tomatoes and getting counted as a separate ingredient. Some ingredients are listed along with the brand name like KRAFT Reduced Fat Shredded Mozzarella Cheese, Johnsonville Smoked Sausage, Johnsonville Mild Italian Sausage Links etc and also constitutes to the ingredient list. This variation makes difficult to get the proper ingredient list for the analysis. Extensive work is needed to clean and correct the noisy data so that proper analysis can be carried out. This correction process is not carried out as part of this project.

Certain ingredients like salt or water etc should be avoided from the analysis as those are not the ingredient we are looking for the analysis. We tried to clean such elements during ingredient relationship analysis but we had little success as those ingredients are present in the dataset in various forms.

4.4 Analysis and Findings

4.4.1 Recipe Distribution By Cuisine. We first analyze entire dataset to understand the total number of recipes and their distribution across various cuisines. We use Pythons Panda library to get the total recipe count as 39774 and plot the distribution. Figure 4 shows number of recipes per cuisine. Dataset is heavily dominated by Italian cuisine followed by Mexican cuisine and with very fewer recipes from Russian and Brazilian cuisines. This also highlights another shortcoming of the dataset that it doesn't have equal representation of all cuisines which might give us biased analysis.

[Figure 4 about here.]

Table1 describes recipe count for every cuisine.

[Table 1 about here.]

4.4.2 Most Used Ingredients All Cuisines. The second analysis is carried out to understand top 20 ingredients getting used across cuisine or globally. Ingredient *Salt* is obvious topper followed by *Oil* and *Onions*. This also proves our craving for saltiness and fat. Figure 5 shows top 20 ingredient across cuisines.

[Figure 5 about here.]

4.4.3 Ingredients Distribution By Cuisines. The third analysis is carried out to understand key ingredient for each cuisine. These key ingredients define those cuisines and provide unique test characterized by that cuisine. We limited ingredient list to top 10 to get the key ingredients for each cuisine. Figure 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 shows top 10 key ingredient used in the corresponding cuisines.

[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.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

[Figure 18 about here.]

[Figure 19 about here.]

[Figure 20 about here.]

[Figure 21 about here.]

[Figure 22 about here.]

[Figure 23 about here.]

[Figure 24 about here.]

[Figure 25 about here.]

4.4.4 Ingredients Relationship. Forth analysis is carried out to understand the relationship between the ingredient to find out ingredient clusters. This analysis helps us understand the ingredient combinations which can be used together to provide great dish every time. This model can be used to predict ingredients for certain recipe based on the cluster. We used Gephi tool to analyze and produce the graph for this analysis. Gephi accepts network structure in terms of Node and Edge relationship. We created this network using python by relating all ingredients present in the recipe with each other. Ingredients become the node and source and target nodes become the edges. These network files generated in excel spreadsheet are converted to CSV format and imported into the Gephi tool. Import created 5405 Nodes and 290828 edges for processing and analysis. Force Atlas 2 layout present in Gephi has been applied to the network which brings nodes with higher weights and shared connections closer to each other. We also used Gephi Data Laboratory to clean up duplicate or unwanted nodes. Filtering based on Degree Range and Edge Weight has been applied to data to reduce node and edges to get the graph which can be used for analysis and avoid crashing Gephi due to large data. Modularity statistic uncovered 5 ingredient clusters which can be identified by different color in the graph. This cluster can approximately related to the cuisines present in our dataset and confirms our earlier analysis of ingredient by cuisine.

- Orange - Mexican
- Brown - Indian
- Blue - Chinese
- Green - Italian
- Gray - Southern US

Figure 26 shows ingredient cluster of more than 1000 nodes.

[Figure 26 about here.]

Figure 27 shows ingredient cluster of around 100 nodes.

[Figure 27 about here.]

4.5 Shortcomings

Improper documentation of ingredient names in the dataset reduces the correctness of this analysis. In absence of proper ingredient name and duplication of ingredient name prevents getting exact ingredient weight into the analysis. A dataset with uniform ingredient name can help this analysis to achieve its best. If we don't find proper ingredient name then this analysis needs to include extensive data cleaning process which can be considered an improvement to this project.

Network file creation algorithm can be enhanced further by considering the number of recipes for the ingredient to provide additional weight to the relationship which can provide the stronger bond between the ingredients.

4.6 Future Work

This dataset can be analyzed to find out ingredient overlap between various cuisine and can provide insight into the influence of one cuisine on another. Usually, geographically neighboring cuisines are influenced by each other as they share common ingredients.

5 CONCLUSION

This project shows most used ingredient, ingredient distribution by cuisine and predictive ingredient relationship model as per the goal of the project. We also show various opportunities present with ingredient data analysis and role of big data analytics. We prove human craving for salty and fatty food as salt and oil are most used ingredient across cuisines as per the analysis. We understand now based on our analysis key ingredient of any cuisine. Ingredient cluster shows why those ingredients are the base of certain cuisine and recipe of those ingredients always turn out delicious. We also crave for the good data so that we can provide more accurate analysis of the ingredients. Ingredient analysis has potential not only to help restaurant and food industry but it can help with our social responsibility of sustainability and understanding different cuisines and culture. As food industries interest grows in big data analytics, we will continue to see more evaluations of the ingredients.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions in this project. The author would also like

to thank Kaggle application for hosting ingredient dataset which is used in this project and various online resource which helped understand Python and Gephi.

REFERENCES

[1] Bagrow James P. Barabsi Albert-Lszl Ahn Yong-Yeol, Ahnert Sebastian E. 2011. Flavor network and the principles of food pairing. *Scientific Reports* 1, 196 (2011). https://www.nature.com/articles/srep00196#supplementary-information

[2] businessdictionary. 2017. Food. web. (2017). http://www.businessdictionary.com/definition/food.html

[3] Usashi Chatterjee, Vinit Kumar, and Devika P. Madalli. 2016. Formalizing Food Ingredients for Data Analysis and Knowledge Organization. 10 (07 2016), 289–309.

[4] S. M. Church. 2015. The importance of food composition data in recipe analysis. web. (2015). http://onlinelibrary.wiley.com/doi/10.1111/nbu.12125/abstract

[5] collinsdictionary. 2017. Recipe. web. (2017). https://www.collinsdictionary.com/us/dictionary/english/recipe

[6] inkhorn82. 2014. A Delicious Analysis. web. (2014). https://www.r-bloggers.com/a-delicious-analysis-aka-topic-modelling-using-recipes/

[7] kaggle. 2015. What's Cooking? web. (2015). https://www.kaggle.com/c/whats-cooking/data

[8] Bernard Lahousse. 2016. Using Big Data to Transform Unfamiliar Ingredients Into Tasty Recipes. web. (2016). https://foodtechconnect.com/2016/04/20/big-food-data-recipes-from-unfamiliar-ingredients/

[9] oxforddictionaries. 2017. Ingredient. web. (2017). https://en.oxforddictionaries.com/definition/ingredient

[10] Matthew Robinson. 2015. Big Data Analytics and Food Come Together At Flavourspace. web. (2015). http://www.theculinaryexchange.com/food-innovation/big-data-analytics-and-food-come-together-at-flavourspace/

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

DONE:

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

DONE:

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

DONE:

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

DONE:

Use a space after . , :

DONE:

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

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

A.5 Citation Issues and Plagiarism

DONE:

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

DONE:

Claims made without citations provided

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material

A.6 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use "quotes" , instead of " "

DONE:

To emphasize a word, use *emphasize* and not "quote"

DONE:

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

DONE:

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

DONE:

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

A.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

Do not artificially in?ate your paper if you are below the page limit

A.8 Details about the Figures and Tables

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

Do not use textwidth as a parameter for includegraphics

DONE:

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

e.g.

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

re

5

List of Figures

1	Flowchart of the Methodology to Analyze Ingredients	7
2	Ingredient Data Structure	7
3	Data Loading	7
4	Recipe Distribution By Cuisine	8
5	Top 20 Ingredients	9
6	Top 10 Ingredients	10
7	Top 10 Ingredients	11
8	Top 10 Ingredients	12
9	Top 10 Ingredients	13
10	Top 10 Ingredients	14
11	Top 10 Ingredients	15
12	Top 10 Ingredients	16
13	Top 10 Ingredients	17
14	Top 10 Ingredients	18
15	Top 10 Ingredients	19
16	Top 10 Ingredients	20
17	Top 10 Ingredients	21
18	Top 10 Ingredients	22
19	Top 10 Ingredients	23
20	Top 10 Ingredients	24
21	Top 10 Ingredients	25
22	Top 10 Ingredients	26
23	Top 10 Ingredients	27
24	Top 10 Ingredients	28
25	Top 10 Ingredients	29
26	Ingredient Cluster	30
27	ingredient Cluster 100 Nodes	31

Figure 1: Flowchart of the Methodology to Analyze Ingredients

```
{
 "id": 24717,
 "cuisine": "indian",
 "ingredients": [
     "tumeric",
     "vegetable stock",
     "tomatoes",
     "garam masala",
     "naan",
     "red lentils",
     "red chili peppers",
     "onions",
     "spinach",
     "sweet potatoes"
 ]
 },
```

Figure 2: Ingredient Data Structure

```
#read the ingredient data using pandas
dfTrain = pd.read_json('./data/train.json')


#load data using json
dataFilePath="./data/train.json"
with open(dataFilePath) as data_file:
    data = json.load(data_file)
```

Figure 3: Data Loading

Figure 4: Recipe Distribution By Cuisine

8

Figure 5: Top 20 Ingredients

[9]

Figure 6: Top 10 Ingredients

10

Figure 7: Top 10 Ingredients

11

Figure 8: Top 10 Ingredients

12

Figure 9: Top 10 Ingredients

13

Figure 10: Top 10 Ingredients

Figure 11: Top 10 Ingredients

15

Figure 12: Top 10 Ingredients

16

Figure 13: Top 10 Ingredients

[17]

Figure 14: Top 10 Ingredients

18

Figure 15: Top 10 Ingredients

19

Figure 16: Top 10 Ingredients

[20]

Figure 17: Top 10 Ingredients

21

Figure 18: Top 10 Ingredients

22

Figure 19: Top 10 Ingredients

23

Figure 20: Top 10 Ingredients

Figure 21: Top 10 Ingredients

Figure 22: Top 10 Ingredients

26

Figure 23: Top 10 Ingredients

27

Figure 24: Top 10 Ingredients

28

Figure 25: Top 10 Ingredients

29

Figure 26: Ingredient Cluster

Figure 27: ingredient Cluster 100 Nodes

List of Tables

1 Recipe Count By Cuisine 33

Table 1: Recipe Count By Cuisine

Cuisine	Recipe Count
brazilian	467
british	804
cajun creole	1546
chinese	2673
filipino	755
french	2646
greek	1175
indian	3003
irish	667
italian	7838
jamaican	526
japanese	1423
korean	830
mexican	6438
moroccan	821
russian	489
southern us	4320
spanish	989
thai	1539
vietnamese	825

```
bibtext report
==========================================

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Warning--page numbers missing in both pages and numpages fields in Ahn2011
Warning--no journal in Chatterjee2016
(There were 16 error messages)
make[2]: *** [bibtex] Error 2

latex report
```

```
=============================================

        [2017-12-04 22.01.24] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.5s.
../README.yml
  48:74     error    trailing spaces  (trailing-spaces)
  49:71     error    trailing spaces  (trailing-spaces)
  50:72     error    trailing spaces  (trailing-spaces)


================================================================================
Compliance Report
================================================================================



name: Sushant Athaley
hid:  302
paper1: Nov 3 2017 100%
paper2: 100%
project: 100%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

33
wc 302 project 33 2935 report.tex
wc 302 project 33 3835 report.pdf
wc 302 project 33 361 report.bib

find "
--------------------------------------------------------------------------------
```

```
80: "id": 24717,

81: "cuisine": "indian",

82: "ingredients": [

83: "tumeric",

84: "vegetable stock",

85: "tomatoes",

86: "garam masala",

87: "naan",

88: "red lentils",

89: "red chili peppers",

90: "onions",

91: "spinach",

92: "sweet potatoes"

106: dataFilePath="./data/train.json"

passed: False

find footnote
--------------------------------------------------------------------------------

passed: True

find input{format/i523}
--------------------------------------------------------------------------------

4: \input{format/i523}

passed: True

find input{format/final}
--------------------------------------------------------------------------------

passed: False
```

```
floats
--------------------------------------------------------------------------------

66: Figure \ref{f:methodology} shows methodology used for this project
    to analyze ingredient data.
67: \begin{figure}[!ht]
68: \centering\includegraphics[width=\columnwidth]{images/methodology.
    PNG}
69: \caption{Flowchart of the Methodology to Analyze Ingredients
    }\label{f:methodology}
76: The dataset for this study is sourced from Kaggle application
    \cite{www-kaggle}. This dataset is publicly available and featured
    in ``What's Cooking?'' competition. This dataset is in JSON format
    and of 12MB size. This dataset contains recipe id, cuisine and
    list of ingredients as described in Figure \ref{c:data-structure}.
77: \begin{figure}[htb]
96: \caption{Ingredient Data Structure}\label{c:data-structure}
98: This dataset contains total 39774 recipes across various cuisines.
    We used two different methods to load this data. Cuisine and
    ingredient analysis is done by loading data into \emph{pandas
    dataframe} and to analyze ingredient relationship data has been
    loaded into \emph{json} object. Figure \ref{c:data-loading} shows
    the code for data loading used in this project.
99: \begin{figure}[htb]
110: \caption{Data Loading}\label{c:data-loading}
120: We first analyze entire dataset to understand the total number of
     recipes and their distribution across various cuisines. We use
     Pythons Panda library to get the total recipe count as 39774 and
     plot the distribution. Figure
     \ref{f:Number_of_recipes_by_cuisine} shows number of recipes per
     cuisine. Dataset is heavily dominated by Italian cuisine followed
     by Mexican cuisine and with very fewer recipes from Russian and
     Brazilian cuisines. This also highlights another shortcoming of
     the dataset that it doesn't have equal representation of all
     cuisines which might give us biased analysis.
121: \begin{figure}[!ht]
122: \centering\includegraphics[width=\columnwidth]{images/Number_of_r
     ecipes_by_cuisine.png}
123: \caption{Recipe Distribution By Cuisine
     }\label{f:Number_of_recipes_by_cuisine}
126: Table\ref{t:recipecount} describes recipe count for every
     cuisine.
127: \begin{table}[htb]
130: \label{t:recipecount}
159: The second analysis is carried out to understand top 20
```

```
ingredients getting used across cuisine or globally. Ingredient
\emph{Salt} is obvious topper followed by \emph{Oil} and
\emph{Onions}. This also proves our craving for saltiness and
fat. Figure \ref{f:Ingredient_Distribution} shows top 20
ingredient across cuisines.
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/Ingredient_
Distribution.png}
\caption{Top 20 Ingredients }\label{f:Ingredient_Distribution}
The third analysis is carried out to understand key ingredient
for each cuisine. These key ingredients define those cuisines and
provide unique test characterized by that cuisine. We limited
ingredient list to top 10 to get the key ingredients for each
cuisine. Figure \ref{f:italian_10_most_used_ingredients},
\ref{f:brazilian_10_most_used_ingredients},
\ref{f:british_10_most_used_ingredients},
\ref{f:cajun_creole_10_most_used_ingredients},
\ref{f:chinese_10_most_used_ingredients},
\ref{f:filipino_10_most_used_ingredients},
\ref{f:french_10_most_used_ingredients},
\ref{f:greek_10_most_used_ingredients},
\ref{f:indian_10_most_used_ingredients},
\ref{f:irish_10_most_used_ingredients},
\ref{f:jamaican_10_most_used_ingredients},
\ref{f:japanese_10_most_used_ingredients},
\ref{f:korean_10_most_used_ingredients},
\ref{f:mexican_10_most_used_ingredients},
\ref{f:moroccan_10_most_used_ingredients},
\ref{f:russian_10_most_used_ingredients},
\ref{f:southern_us_10_most_used_ingredients},
\ref{f:spanish_10_most_used_ingredients},
\ref{f:thai_10_most_used_ingredients},
\ref{f:vietnamese_10_most_used_ingredients}  shows top 10 key
ingredient used in the corresponding cuisines.
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/italian_10_
most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:italian_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/brazilian_1
0_most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:brazilian_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/british_10_
```

```
most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:british_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/cajun_creol
e_10_most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:cajun_creole_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/chinese_10_
most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:chinese_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/filipino_10
_most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:filipino_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/french_10_m
ost_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:french_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/greek_10_mo
st_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:greek_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/indian_10_m
ost_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:indian_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/irish_10_mo
st_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:irish_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/jamaican_10
_most_used_ingredients.png}
\caption{Top 10 Ingredients
}\label{f:jamaican_10_most_used_ingredients}
\begin{figure}[!ht]
\centering\includegraphics[width=\columnwidth]{images/japanese_10
_most_used_ingredients.png}
```

```
224: \caption{Top 10 Ingredients
     }\label{f:japanese_10_most_used_ingredients}
227: \begin{figure}[!ht]
228: \centering\includegraphics[width=\columnwidth]{images/korean_10_m
     ost_used_ingredients.png}
229: \caption{Top 10 Ingredients
     }\label{f:korean_10_most_used_ingredients}
232: \begin{figure}[!ht]
233: \centering\includegraphics[width=\columnwidth]{images/mexican_10_
     most_used_ingredients.png}
234: \caption{Top 10 Ingredients
     }\label{f:mexican_10_most_used_ingredients}
237: \begin{figure}[!ht]
238: \centering\includegraphics[width=\columnwidth]{images/moroccan_10
     _most_used_ingredients.png}
239: \caption{Top 10 Ingredients
     }\label{f:moroccan_10_most_used_ingredients}
242: \begin{figure}[!ht]
243: \centering\includegraphics[width=\columnwidth]{images/russian_10_
     most_used_ingredients.png}
244: \caption{Top 10 Ingredients
     }\label{f:russian_10_most_used_ingredients}
247: \begin{figure}[!ht]
248: \centering\includegraphics[width=\columnwidth]{images/southern_us
     _10_most_used_ingredients.png}
249: \caption{Top 10 Ingredients
     }\label{f:southern_us_10_most_used_ingredients}
252: \begin{figure}[!ht]
253: \centering\includegraphics[width=\columnwidth]{images/spanish_10_
     most_used_ingredients.png}
254: \caption{Top 10 Ingredients
     }\label{f:spanish_10_most_used_ingredients}
257: \begin{figure}[!ht]
258: \centering\includegraphics[width=\columnwidth]{images/thai_10_mos
     t_used_ingredients.png}
259: \caption{Top 10 Ingredients
     }\label{f:thai_10_most_used_ingredients}
262: \begin{figure}[!ht]
263: \centering\includegraphics[width=\columnwidth]{images/vietnamese_
     10_most_used_ingredients.png}
264: \caption{Top 10 Ingredients
     }\label{f:vietnamese_10_most_used_ingredients}
278: Figure \ref{f:ingredient_modularity} shows ingredient cluster of
     more than 1000 nodes.
279: \begin{figure}[!ht]
280: \centering\includegraphics[width=\columnwidth]{images/ingredient_
```

```
     modularity.png}
281: \caption{Ingredient Cluster }\label{f:ingredient_modularity}
284: Figure \ref{f:ingredient_modularity100} shows ingredient cluster
     of around 100 nodes.
285: \begin{figure}[!ht]
286: \centering\includegraphics[width=\columnwidth]{images/ingredient_
     modularity100.png}
287: \caption{ingredient Cluster 100 Nodes
     }\label{f:ingredient_modularity100}

figures 27
tables 1
includegraphics 25
labels 28
refs 9
floats 28

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

Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors
```

```
bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6169 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6261 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Ahn Yong-Yeol, Ahnert Sebastian E., Bagrow James P., Barab
while executing---line 6457 of file ACM-Reference-Format.bst
Warning--page numbers missing in both pages and numpages fields in Ahn2011
Warning--no journal in Chatterjee2016
(There were 16 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
```

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. They 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 statics 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 Austrailia (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 patientially initiall 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 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 following 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 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 options for autonomous doctor's visits or telemedicine 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 telemedicine[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 patientfis 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 Unites 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 on 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, martial 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 a 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 it 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] Malarie Gokey. 2016. Why smart clothes, not watches, are the future of wearables. Online. (01 2016). https://www.digitaltrends.com/wearables/smart-clothing-is-the-future-of-wearables/

[21] World Bank Group. [n. d.]. Health Expenditure per Capita (current US$). Online. ([n. d.]). https://data.worldbank.org/indicator/SH.XPD.PCAP?end=2014&name_desc=true&start=2004&view=chart

[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

[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://qpp.cms.gov/apms/overview

[40] United States Department of Health and Human Services. 2017. Health IT Dashboard. Online. (08 2017). https://dashboard.healthit.gov/quickstats/quickstats.php

[41] United States Department of Labor. [n. d.]. Long-Term Price Trends for Computers, TVs, and Related Items. Online. ([n. d.]). https://www.bls.gov/opub/ted/2015/long-term-price-trends-for-computers-tvs-and-related-items.htm

[42] Optum. [n. d.]. Data Rich, Insight Poor. Online. ([n. d.]). https://cdn-aem.optum.com/content/dam/optum3/optum/en/images/infographics/Game_changer_Track_Two_04_Data_Rich_Insight_Poor_Infog_Images_2016.pdf

[43] World Health Organizaion. [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/

[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/

[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

[47] World Healh 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/

[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
[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
===========================================

        [2017-12-04 22.01.50] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
```

```
Typesetting of "report.tex" completed in 1.0s.

================================================================================
Compliance Report
================================================================================


name: Marks, Paul
hid:  327
paper1: 100%  10/25/2017
paper2: 100%  11/06/17
project: 99%

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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------


bibtex
--------------------------------------------------------------------------------


label errors


bibtex errors

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

```
Database file #1: report.bib
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
```

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

Sean M. Shiverick
Indiana University Bloomington
smshiver@indiana.edu

ABSTRACT

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

KEYWORDS

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

1 INTRODUCTION

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

1.1 Opioid Overdose Epidemic

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

1.2 Machine Learning Approaches

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

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

1.3 Classification Algorithms

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

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

1.3.2 Tree Based Models. Decision tree models are widely used for classification and regression. Tree models "learn" a hierarchy of if-else questions that are represented in the form of a decision tree. Building decision trees proceeds from a root node as the starting point and continues through a series of decisions or choices. Each node in the tree either represents either a question or a terminal node (i.e.,leaf) that contains the outcome. Applied to a binary classification task, the decision tree algorithm *learns* the sequence of if-else questions that arrives at the outcome most quickly. For data with continuous features, the decisions are expressed in the form of, "Is feature x larger than value y?" [10] In constructing the tree, the algorithm searches through all possible decisions or tests, and find a solution that is most informative about the target outcome. A decision tree classifier is used for binary or categorical targets, and decision tree regression is used for continuous target outcomes. The recursive branching process of tree based models yields a binary tree of decisions, with each node representing a test that considers a single feature. This process of recursive partitioning is repeated until each leaf in the decision tree contains only a single target. Prediction for a new data point proceeds by checking which region of the partition the point falls in, and predicting the majority in that feature space. The main advantage of tree based models is that they require little adjustment and are easy to interpret. A drawback is that they can lead to very complex models that are highly overfit to the training data. A common strategy to prevent overfitting is *pre-pruning*, which stops tree construction early by limiting the maximum depth of the tree, or the maximum number of leaves. One can also set the minimum number of points in a node required for splitting. Another approach is to build the tree and then remove or collapse nodes with little information, which is called *post-pruning*. Decision trees work well with features measured on very different scales, or with data that has a mix of binary and continuous features.

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

1.4 Project Goals

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

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

2 METHOD

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

2.1 Data

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

2.2 Data Cleaning and Preparation

2.2.1 Data Cleaning. All steps of this analysis was completed in a python interactive notebook [16] based following examples from *Python for Data Analysis* [9]. After saving the NSDUH-2015 as a data frame object, the dataset was subset by columns to include demographic characteristics (e.g., age category, sex, marital status, education, employment status, and category of metropolitan area), measures of physical health (e.g., overall health, STDs, Hepatitis, HIV, Cancer, hospitalization), mental health (e.g., Adult Depression, Emotional Distress, Suicidal Thoughts, Plans), Suicide Attempts, Pain Reliever Medication Use, Misuse, and Abuse (over past year, past month), Prescription Opioid Medications Taken in Past year (e.g., Hydrocodone, Oxycodone, Tramadol, Morphine, Fentanyl, Oxymorphone, Demerol, Hydromorphone), Heroin Use, Abuse (over past year, past month), Tranquilizer Use, Sedative Use, Cocaine Use, Amphetamine and Methamphetamine Use, Hallucinogen Use, Drug Treatment (e.g., Inpatient, Outpatient, Hospital, Mental Health Clinic, ER, Drug Treatment Status), and Mental Health Treatment History. A codebook was created to provide a complete list of variables included with summaries of response categories [19]. The following steps were taken to detect and remove inconsistencies in the data [13]:

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

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

3 RESULTS

3.1 Exploratory Data Analysis

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

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

[Table 1 about here.]

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

[Table 2 about here.]

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

[Figure 1 about here.]

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

[Figure 2 about here.]

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

[Figure 3 about here.]

3.2 Classifier Models of Heroin Use

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

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

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

[Figure 4 about here.]

3.2.2 Decision Tree Classifier. The following analysis used the *Decision Tree Classifier* package in Scikit-Learn, which only does pre-pruning. First, the decision model was build using the default setting of a fully developed tree until all leaves are pure. The random state' features is fixed to break ties internally. Accuracy on the training set was 0.99 and test set accuracy was 0.974. Without restricting their depth, decision trees can become complex; unpruned trees are prone to overfitting and do not generalize well to new data. Limiting the depth of tree decreases overfitting, which results in lower training set accuracy, but improved performance on the test set. Next, pre-pruning was applied, with a maximum depth of 4, which means the algorithm split on four consecutive questions. Training set accuracy of the pruned tree was 0.985 and test set accuracy was 0.984. Even with a depth of 4, the tree can become a bit complex. Figure 5 shows a partial view of the decision tree classifier of heroin use (the entire tree was too wide to include as a legible Figure), and the full tree image is available in the notebook BDA-Analytics-Classifier-Heroin.ipynb [17]. The decision tree shows the top features that the algorithm split on to classify heroin use. One way to interpret a decision tree it by following the sample numbers represented at the test split for each node. The classifier algorithm selected Cocaine Use (aggregated score) as the root node of the decision tree. The branch to the left side of the tree represents samples with a score equal to or less than 1.5 (n=40956), whereas the branch to the right represents samples with a Cocaine Use score greater than 1.5 (n=1903). The second split on the right occurs for Any Prescription Opioid Pain Reliever Use (PRLANY), with n=1443 having a score less than or equal to 3.5, and n=460 respondents with a PRL score greater than 3.5. In other words, of those respondents who reported relatively high Cocaine use, a small portion also reported relatively high Prescription Opioid PRL use. Instead of looking at the whole tree, features importance is a common summary function that rates how important each feature is for the classification decisions made in the algorithm. Each feature is assigned an importance value between 0 and 1; with a value of 1 indicating the feature perfectly predicts the target and a value of 0 meaning that the feature was not used at all. Feature importance values also always sum to 1. A feature may have a low feature importance value because another feature encodes the same information. The top two important features for classifying Heroin Use were Cocaine Use and Any Prescription Opioid PRL Use, with smaller importance given to Opioid PRL Misuse Ever and Prescription Opioid PRL Misuse and Abuse.

[Figure 5 about here.]

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

[Figure 6 about here.]

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

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

[Figure 7 about here.]

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

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

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

[Figure 8 about here.]

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

[Figure 9 about here.]

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

[Figure 10 about here.]

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

[Figure 11 about here.]

4 DISCUSSION

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

4.1 Comparison of Classifier Models

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

4.2 Study Limitations

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

4.3 Extension to Big Data

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

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

4.4 Opioid Addiction and Epidemic Spreading

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

5 CONCLUSION

This project compared several classification algorithms to predict heroin use and prescription opioid misuse and abuse. The results provided partial support for the hypothesis that prescription opioid misuse is associated with the use of illicit opioids such as heroin. Several features were identified as important for classifying heroin use, including Cocaine Use, Amphetamine Use, and any prescription opioid medication use. In regards to predicting heroin use, it appears the use of other illicit drugs such as Cocaine and Amphetamine was perhaps more informative than any prescription opioid use or misuse. Heroin use was selected as important for classifying prescription opioid pain reliever misuse, but additional factors also played as role, including tranquilizer use, age category, overall health, cocaine use. Substance treatment had the largest regression coefficient, suggesting that people who are misusing prescription opioid pain medication are also more likely to be in drug treatment programs. The direction of these effects cannot be determined owing to the nature of the analyses. On the one hand individual misusing or abusing prescription opioids may also be using heroin. Alternatively, individuals with a susceptability for opioid use may be equally likely to have use heroin and also to have misused prescription opioids. A general conclusion is that there is that persons who report misusing prescription opioids were twice as likely to have used heroin. The results do not provide sufficient evidence to rule out the alternative hypothesis. Given the relatively low rates of opioid and heroin in this sample, additional evidence is needed to resolve this question. The study can provide information to raise awareness about the risk factors for prescription opioid addiction and may help reduce opioid overdose deaths.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Substance Abuse, Center for Behavioral Health Statistics Mental Health Services Administration, and Quality. 2016. *National Survey on Drug Use and Health (NSDUH) 2015.* Online data archive. United States Department of Health and Human Services., Ann Arbor, MI. https://doi.org/10.3886/ICPSR50011.v1

[2] Vittoria Colizza, Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7 (2006), 2015–2020. https://doi.org/10.1073/pnas.0510525103 arXiv:http://www.pnas.org/content/103/7/2015.full.pdf

[3] Centers for Disease Control and Prevention. 2017. Prescription Opioid Overdose Data. online. (Oct. 2017). https://www.cdc.gov/drugoverdose/data/overdose.html

[4] hd1 and yoavram. 2016. Python: Download Returned Zip file from URL. Online. (Feb. 2016). https://stackoverflow.com/questions/9419162/python-download-returned-zip-file-from-url Stackoverflow.com.

[5] M. Herland, T. M. Khoshgoftaar, and R. Wald. 2014. A review of data mining using big data in health informatics. *Journal Of Big Data* 1, 2 (2014). https://doi.org/10.1186/2196-1115-1-2

[6] K. Johnson, A. Isham, D.V. Shah, and D.H. Gustafson. 2011. Potential Roles for New Communication Technologies in Treatment of Addiction. *Current psychiatry reports.* (2011). https://doi.org/10.1007/s11920-011-0218-y

[7] Rose A. Judd, Noah Aleshire, Jon E. Zibbell, and R. Matthew Gladden. 2016. *Increases in Drug and Opioid Overdose Deaths, United States, 2000-2014.* techreport 64(50). Centers for Disease Control and Prevention, Atlanta, GA. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm Morbidity and Mortality Weekly Report (MMWR).

[8] Lisa A. Marsch. 2012. Leveraging teachnology to enhance addiction treatment and recovery. *Journal of Addictive Diseases* 31, 3 (2012), 313–318. https://doi.org/10.1080/10550887.2012.694606

[9] Wes McKinney. 2017. *Python for Data Analysis.* O'Reilly Media Inc., Sebastopol, CA. https://github.com/wesm/pydata-book

[10] Andreas C. Muller and Sarah Guido. 2017. *Introduction to Machine Learning.* O'Reilly, Sebastopol, CA. https://github.com/amueller/introduction_to_ml_with_python/

[11] National Institute on Drug Abuse (NIDA). 2017. *Overdose Death Rates.* Summary. National Institutes of Health (NIH), Washington D.C. https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

[12] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* 86 (Apr 2001), 3200–3203. Issue 14. https://doi.org/10.1103/PhysRevLett.86.3200

[13] E. Rahm and H. Hai Do. 2000. *Data cleaning: Problems and current approaches.* techreport 23(4). Bulletin of the Technical Committee on Data Engineering, 1730 Massachusetts Avenue, Washington D.C. https://s3.amazonaws.com/academia.edu.documents/41858217/A00DEC-CD.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511155930&Signature=VWRM7u4KWtP6ZxX5jB%2Bh6wMCbpg%3D&response-content-disposition=inline%3B%20filename%3DAutomatically_extracting_structure_from.pdf#page=5

[14] Sebastian Raschka and Vahid Mirjalili. 2017. *Ptyhon Machine Learning, Second Edition*. Packt, Birmingham, UK. https://github.com/rasbt/python-machine-learning-book-2nd-edition
[15] Yavin Shaham, Uri Shalev, Lin Lu, Harriet de Wit, and Jane Stewart. 2003. The reinstatement model of drug relapse: history, methodology and major findings. *Psychopharmacology* 168, 1 (01 Jul 2003), 3–20. https://doi.org/10.1007/s00213-002-1224-x
[16] S.M. Shiverick. 2017. BDA Project Data. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Data.ipynb
[17] S.M. Shiverick. 2017. Classification Models of Heroin Use. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-Heroin.ipynb Interactive Python Jupyter Notebook.
[18] S.M. Shiverick. 2017. Classification Models of Prescription Opioid Pain Relievers Misuse. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-PRL.ipynb Interactive Python Jupyter Notebook.
[19] S.M. Shiverick. 2017. Project Codebook for Data Variables from NSDUH-2015. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/project-data-codebook.txt
[20] S. M. Shiverick. 2017. Exploratory Data Analysis. Github. (Dec. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb
[21] S. M. Shiverick. 2017. Project Data Visualization. Github. (Dec. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Explore-Data.ipynb
[22] S. M. Shiverick. 2017. Project Workflow Pipeline. Github. (Dec. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/readme.md
[23] J. Swendsen. 2016. Contributions of mobile technologies to addiction research. *Dialogues Clinical Neuroscience* 18, 2 (June 2016), 213–221. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969708/
[24] Jake VanderPlas. 2017. *Python Data Science Handbook*. O'Reilly Media Inc., Sebastopol, CA. https://jakevdp.github.io/PythonDataScienceHandbook/
[25] Upkar Varshney. 2013. Smart medication management system and multiple interventions for medication adherence. *Decision Support Systems* 55, 5 (May 2013), 538–551. https://doi.org/10.1016/j.dss.2012.10.011
[26] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. 2014. Medication-Assisted Therapies: Tackling the Opioid-Overdose Epidemic. *New England Journal of Medicine* 370, 22 (2014), 2063–2066. https://doi.org/10.1056/NEJMp1402780 arXiv:http://dx.doi.org/10.1056/NEJMp1402780 PMID: 24758595.

A CODE REFERENCES

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

A.1 Download and Extract Data

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

A.2 Data Cleaning and Preparation

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

A.3 Exploratory Data Analysis

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

A.4 Data Visualization

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

A.5 Classification Algorithms

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

B ISSUES

DONE:

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

B.1 Assignment Submission Issues

DONE:

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

B.2 Uncaught Bibliography Errors

DONE:

Missing bibliography file generated by JabRef

DONE:

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

DONE:

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

B.3 Formatting

DONE:

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

DONE:

Other formatting issues

B.4 Writing Errors

DONE:

Errors in title, e.g. capitalization

DONE:

Spelling errors

DONE:

Are you using *a* and *the* properly?

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

Use a space after . , :

DONE:

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

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

B.5 Citation Issues and Plagiarism

DONE:

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

DONE:

Claims made without citations provided

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material

B.6 Character Errors

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

B.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

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

B.8 Details about the Figures and Tables

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

Do not use textwidth as a parameter for includegraphics

DONE:

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

e.g.

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

re

List of Figures

1 Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin 13

2 Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size 14

3 Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area 15

4 Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C) 16

5 Decision Tree Classification of Heroin Use (Partial View) 17

6 Feature Importance for Random Forests Classifier for Heroin Use 18

7 Feature Importance for Gradient Boosting Classifier for Heroin Use 19

8 Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty 20

9 Decision Tree for Prescription Opioid (PRL) Misuse 21

10 Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse 22

11 Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse 23

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

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

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

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

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

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

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

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

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

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

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

List of Tables

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

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

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

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

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

```
bibtext report
===========================================

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

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


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


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


latex report
===========================================

        [2017-12-04 22.02.33] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Label `tab:freq' multiply defined.
p.8   L929  : [granovetter73] undefined
p.8   L933  : [watts98] undefined
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
There were undefined citations.
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
```

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

================================================================================
Compliance Report
================================================================================


name: Sean Shiverick
hid:  335
paper1: 10/25/17 100%
paper2: 100%
project: 100%

yamlcheck
--------------------------------------------------------------------------------


wordcount
--------------------------------------------------------------------------------

25
wc 335 project 25 8439 report.tex
wc 335 project 25 9531 report.pdf
wc 335 project 25 1078 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
--------------------------------------------------------------------------------

352: \begin{table}
355: \label{tab:freq}
400: \begin{table}
403: \label{tab:freq}
432: \begin{figure}[!ht]
433: \centering\includegraphics[width=\columnwidth]{images/Figure1.pdf
     }
436: \label{f:Figure1}
460: \begin{figure}[!ht]
461: \centering\includegraphics[width=\columnwidth]{images/Figure2.pdf
     }
464: \label{f:Figure2}
499: \begin{figure}[!ht]
500: \centering\includegraphics[width=\columnwidth]{images/Figure3.pdf
     }
503: \label{f:Figure3}
549: \begin{figure}[!ht]
550: \centering\includegraphics[width=\columnwidth]{images/Figure4.pdf
     }
553: \label{f:Figure4}
598: \begin{figure}[!ht]
599: \centering\includegraphics[width=\columnwidth]{images/Figure5.pdf
     }
601: \label{f:Figure5}
636: \begin{figure}[!ht]
637: \centering\includegraphics[width=\columnwidth]{images/Figure6.pdf
     }
639: \label{f:Figure6}
673: \begin{figure}[!ht]
674: \centering\includegraphics[width=\columnwidth]{images/Figure7.pdf
     }
676: \label{f:Figure7}
713: \begin{figure}[!ht]
714: \centering\includegraphics[width=\columnwidth]{images/Figure8.pdf
     }
718: \label{f:Figure8}
744: \begin{figure}[!ht]
745: \centering\includegraphics[width=\columnwidth]{images/Figure9.pdf
     }
747: \label{f:Figure9}
```

```
765: \begin{figure}[!ht]
766: \centering\includegraphics[width=\columnwidth]{images/Figure10.pd
     f}
769: \label{f:Figure10}
794: \begin{figure}[!ht]
795: \centering\includegraphics[width=\columnwidth]{images/Figure11.pd
     f}
798: \label{f:Figure11}

figures 11
tables 2
includegraphics 11
labels 13
refs 0
floats 13

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

Label/ref check
89: abusing prescribed opioid medication who also used heroin, shown
    in Figure 1.
332: 386 females). Table 1 shows the raw counts of individual
     substance use by age
391: Table 2 shows the frequency of individuals reporting that they
     had experienced
420: Figure 1 shows the proportion of individuals who reported
     misusing prescription
422: Figure 1 shows the majority of respondents (89 percent) stated
     they had never
425: Figure 1 shows that, of those individuals who reported using
     heroin, the
440: Figure 2 shows the aggregated measure of Opioid Pain Reliever
     misuse and abuse
468: Figure 3 shows the pairplots of demographic features including
     mental health
534: few features. Figure 4 plots the coefficients of logistic
     regression classifier
540: accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows
     that the
571: Figure 5 shows a partial view of the decision tree classifier of
     heroin use
624: feature importance for a single tree. Figure 6 shows the feature
     importance
```

```
664: or by reducing the learning rate. Figure 7 shows that the feature
     importance
702: Figure 8 plots the coefficients of logistic regression classifier
     for
704: Figure 8, the features with the highest coefficient values were
     Treatment
727: of the pruned tree was 0.902 and test set accuracy was 0.902.
     Figure 9 shows
730: notebook) \cite{classifyPRL}. As Figure 9 shows, the decision
     tree classifier
755: 0.896, which suggests that the model overfit the data. Figure 10
     shows the
755: 0.896, which suggests that the model overfit the data. Figure 10
     shows the
757: PRL Misuse. As Figure 10 shows, several features were identified
     as important
757: PRL Misuse. As Figure 10 shows, several features were identified
     as important
783: prescription opioid abuse. Figure 11 shows the feature importance
     for the
783: prescription opioid abuse. Figure 11 shows the feature importance
     for the
784: gradient boosting classifier tree. As Figure 11 shows, several
     features were
784: gradient boosting classifier tree. As Figure 11 shows, several
     features were
817: used heroin than had not (see Figure 1), which partially supports
     the
passed: False -> labels or refs used wrong

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



WARNING: algorithm and below may be used improperly
```

```
126: classification algorithms are considered below.

bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

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

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



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

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

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

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

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

passed: True
```

Comparison between different classification algorithms in Digit Recognizer

Junjie Lu
Indiana University Bloomington
3322 John Hinkle Place
Bloomington, Indiana 47408
junjlu@iu.edu

Yuchen Liu
Indiana University Bloomington
1750 N Range Rd
Bloomington, Indiana 47408
liu477@iu.edu

Wenxuan Han
Indiana University Bloomington
1150 S Clarizz Blvd
Bloomington, Indiana 47401-4294
wenxhan@iu.edu

ABSTRACT

Digit Recognizer is becoming more and more important in many different areas, such as zip code recognizer, banking receipt and balance sheet. Many technology companies are trying to use Big Data to develop more efficient and accurate algorithm for Digit Recognizer. This project uses Digit Recognizer data set from Kaggle.com. There are more than 42000 samples in the data set. Each sample contains 784 features which contain pixel information from a $28 * 28$ graph. Each pixel has a value between 0 to 255. We use binary classification technique for data cleaning and PCA for feature extraction. For the classification model, we choose five most commonly used classification algorithms, which include Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). From the result, SVM classifier on PCA data produces the highest accuracy with 0.9813. The time spend is 127 seconds. Naive Bayes classifier on PCA data spends the least amount of time to finish the classification task. It takes less one second and reaches a 0.8651 accuracy.

KEYWORDS

I523, HID213, HID214, HID209, Big Data, Digit Recognition, Cross Validation, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine

1 INTRODUCTION

People have made a great improvement in digital recognition in recent years. And it plays significant roles in many different areas. Zip code recognizer can scan zip code for post office automatically. Recognizer in banks can help managing user account by scanning their account number. They help people a lot in increasing working efficiency. And many new productions use digital recognition to authenticate password. In this situation, the accuracy and efficiency of recognition become more and more essential and methods in order to increase the accuracy and efficiency are also required.

Fortunately, people have already developed many different types of techniques to avoid faults and decrease running time in recent few years. Several algorithms will be mentioned here. Logistic regression, the most frequently used algorithm in the field of machine learning, also has a good performance in digital recognition. Decision tree is commonly used in decision analysis. It can identify strategies to get a result, in this case, it can also play an important role in digital recognition. Naive Bayes classifier would also be used. Random forest is also a widely used technique in the field of classification and regression. Its special structure with the multitude of decision trees would help it get a fantastic result. Support vector machine can efficiently perform non-linear classification hence it also be considered frequently. These algorithms have different structures so that they have different performance. We can also observe running time and accuracy of different algorithms with different kind of data. In this paper, we are going to talk about this and make the comparison between algorithms in accuracy and efficiency.

2 EXPERIMENT PREPARATION

In this paper, we choose the data of Digit Recognizer from Kaggle.com in order to test different classification algorithms [5]. The goal of this experiment is to correctly identify digits from a data set of tens of thousands of handwritten images. Thus, we could compare the pros and cons of each technique through the recognition accuracy and time-consuming.

2.1 Data Set Description

In train.csv data file, it contains 42000 gray-scale images of hand-drawn digits, from zero through nine. Each image is a 28×28 pixels matrix with a total of 784 pixels [5]. Each pixel has a single pixel-value which is an integer from 0 to 255 associated with it, indicating the lightness or darkness of that pixel (higher numbers meaning lighter). In this experiment, we have plotted the graph in order to see the appearance of these digits easily. Figure 1 shows the first 70 samples.

[Figure 1 about here.]

The training data set has 785 columns. The first column called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. Each pixel column in the training set has a name like $pixelx$, where x is an integer between 0 and 783. To locate a pixel on the image, suppose that we have decomposed x as $x = i * 28 + j$, where i and j are integers between 0 and 27. Then $pixelx$ is located in row i and column j of this matrix [5]. Visually, if we omit the "pixel" prefix, the pixels make up the image like the following form:

$$\begin{matrix} 000 & 001 & 002 & 003 & \cdots & 026 & 027 \\ 028 & 029 & 030 & 031 & \cdots & 054 & 055 \\ 056 & 057 & 058 & 059 & \cdots & 082 & 083 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 728 & 729 & 730 & 731 & \cdots & 754 & 755 \\ 756 & 757 & 758 & 759 & \cdots & 782 & 783 \end{matrix}$$

2.2 Data Cleaning

As we mentioned above, it can be seen from both the figure and the pixel-value that the value varies from 0 to 255, which means each feature is a continuous value. Thus, it is possible that such continuous values might affect our later feature selection. Our observation shows that the values are not very high at the boundaries of 0 and > 0. So here exist three ways to handle it [23]:

(1) Not do any processing on image;
(2) Binarize the image. That is, for the values which are 0, keep them as 0; for the values which are greater than 0, change to 1;
(3) Binarize the image by setting a threshold. That is, for the values which are greater than this threshold, change to 1; otherwise, change to 0.

Obviously, method (2) and (3) will cause the loss of the original information. However, this information may not as important as our expected during the execution of classification algorithms, it could play a positive role in increasing the performance without reducing the accuracy.

In our experiment, we selected method (2) to clean the raw data. The following part of codes shows this operation.

```
from numpy import *

# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0

    print("Data clean completed.")
    return new_data
```

2.3 Feature Extraction

Dimension reduction in the field of machine learning refers to using a mapping method to map the data points in the original high-dimensional space into the low-dimensional space. The essence of dimension reduction is to learn a mapping function $f : x- > y$, where x is the expression of the original data point, y is the low-dimensional vector representation after the data point mapping [9].

The reason why we use data after dimension reduction is that the redundant information and noise information are contained in the original high-dimensional space, which reduces the accuracy of our model. By dimension reduction, we hope to reduce the error caused by redundant information and improve the accuracy of identification. We also hope to find the intrinsic structure of the data structure through the dimension reduction algorithm. Also, in this example, there are 784 features in our data. Space, time and computation complexity are all unacceptable. There are many different dimension reduction algorithms for us to choose. In this project, we choose to use Principle Component Analysis (PCA).

2.3.1 PCA

Principal Component Analysis (PCA) is the most commonly used method of supervised linear dimension reduction. Its goal is to map high-dimensional data to a low-dimensional representation of space by some kind of linear projection. The variance of the data is expected to be maximized in the projected dimension. By keep the variance of data as high as possible, PCA can reduce the dimension of data and keep the loss of information of the data as a minimum [3].

A common understanding is that if all the points are mapped together, almost all information (such as the distance between points) is lost. If the post-mapping variance is as large as possible, the data points are spread apart to preserve more information. It can be proved that PCA is a linear dimension reduction method that loses the original least data information.

One of the questions we faced while we are using PCA is that: how many components should we choose for the model after dimension reduction. In order to solve this problem, we use Explained Variance as our threshold standard. Explained Variance is an important indicator of PCA dimension reduction. The Explained Variance shows the amount of variance explained by each of the selected components. The first column of the PCA model always explains the most variance and the variance explained will keep decrease as the number of column increase. Generally, a dimension with a cumulative contribution rate of about 90% is selected as a reference dimension for PCA dimensionality reduction. In this project, in order to get a more accurate result, we choose 95% as our threshold.

```
from sklearn.decomposition import PCA

def feature_selection(data):
    pca = PCA()
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            break

    # Then, PCA the model by the number of components
    pca = PCA(n_components=n, whiten=True)
    return pca.fit_transform(data)
```

After calculating the explained variance for each component, we decide to choose 30 components for our model. Which shows that there will be 30 features in our model.

3 EXPERIMENT ALGORITHMS

We aim to select five most commonly used classification algorithms which include Decision Tree, Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). This section offers a broad overview of these algorithms before applying them to the digit recognizer problem to compare their characteristics. Then, the result of the different algorithm on different data will show on a table.

For each algorithm, we use:

(1) PCA data - data after using PCA to reduce the dimension on raw data
(2) Clean data - data after our data cleaning process, which set all values greater than 0 to 1 in our data
(3) PCA Clean daata - data after using PCA to reduce the dimension on clean data after data cleaning process

3.1 Cross-Validation

When we build the model, it is normal to follow the principle of simplification since the simpler model we built, the better performance we will get. However, for some complicated problems, our model will also become more complex which might cause the overfitting problem. In order to solve this problem, we introduce the cross-validation technique. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it [13].

The purpose of cross-validation is to select the model with the optimal parameters. After the model is set up, tuning the parameters is a very time-consuming process. Through cross-validation, we can get the model with the optimal parameters much easier. Here are some steps about cross-validation procedure:

(1) Prepare the candidate models, $M_1, M_2, M_3, \cdots$ (the model framework is consistent, only different on the parameters);
(2) For each model, use cross-validation to return the accuracy and error rate information of the model, the result should be the average of cross-validation;
(3) Select the best model by comparing the accuracy or error of the different models.

There are some types of cross-validation which are common to use: K-fold cross-validation and Leave-one-out cross-validation.

- K-fold:
  This method is to divide the data set into k subsets. Each time, select one of the k subsets as the test set and the other $k-1$ subsets become a training set. Then the average accuracy or error across all k trials is computed [13]. In general, we choose 10 as the value of k.
- Leave-one-out (LOO):
  This method is K-fold cross-validation taken to its logical extreme, with $k = n$ ($n > k$), the number of data points in the set [13]. That is, it randomly select n samples as a training set and the rest as a test set. Since the time complexity of this cross-validation is factorial, it is not an appropriate method for big data set.

In this project, we use K-fold cross validation technique to reduce over-fitting of our model and increase the accuracy in each model. We use the function *cross_val_score* from the sklearn package. It have several important parameters to set [7].

(1) CV: int, cross-validation generator or an iterable, optional
This parameter determines the cross-validation splitting strategy, which determined the number of fold we need to use. In our project, we use the default 3-fold cross-validation. Because 3-fold provide us the result in a reasonable time and accuracy.
(2) Scoring: string, callable or None, optional, default: None
The scoring parameter determines what to return after we call the function. We just this parameter to 'accuracy', which will return the accuracy between 0 to 1 for each model.

After we receive the result for each validation, we generate the mean of each result and use the result as the accuracy of the model.

```
from datetime import datetime
from sklearn.cross_validation import cross_val_score

def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label, cv=5,
         scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds

    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)
```

3.2 Decision Tree

3.2.1 Introduction

Decision tree builds classification or regression models in the form of a tree structure (either binary or non-binary) [17]. Each of its non-leaf nodes represents a test on the characteristic attributes, and each leaf node stores a category. The process of decision making using decision tree has the following steps [19]:

(1) Start at the root node;
(2) Test the corresponding characteristic attribute of the items that need to be classified;
(3) Select the branch based on the value until the leaf node is reached;
(4) The category of stored in the leaf node is the result.

The decision tree construction process rely on attribute selection metrics in order to choose the attribute which has the capability to divide tuples into different classes best. The key step in constructing a decision tree is split attributes which means to construct different branches according to the different partition of a certain characteristic attribute at a node. The goal of this step is to make each split subset as "pure" as possible. Split attributes are divided into three different situations:

(1) Attributes are discrete values and do not require to generate a binary decision tree. This time, each partition of an attribute becomes a branch;
(2) Attributes are discrete values and require to generate a binary decision tree. This time, a subset of attribute partitions is used for testing, broken down two branches according

to "subordinate to this subset" and "not subordinate to this subset";

(3) Attributes are continuous values. This time, determine a value as a *split_point* and generate two branches according to > *split_point* and ≤ *split_point*.

There are many attribute selection metric algorithms (e.g. ID3, C4.5, CART, etc.), generally using top-down recursive method with non-backtrack greedy strategy. In our experiment, we applied optimized version of the Classification And Regression Trees (CART) algorithm from scikit-learn library.

The CART algorithm uses a binary recursive segmentation technique [1]: the current sample set is divided into two sub-sample sets, so that each non-leaf node have two branches. Therefore, the decision tree generated by the CART algorithm is a concise binary tree with the root node represents a single input variable (x) and a split point on that variable and the leaf nodes contain an output variable (y) which has the capability to make a prediction [1].

The first key step of CART algorithm is creating the tree model, it examines each variable and all possible partitions of this variable to observe the best partitions. For discrete values such as $U = \{x, y, x\}$, there are three cases of partitions [6]:

$$\{\{x, y\}, \{z\}\}, \{\{x, z\}, \{y\}\}, \{\{y, z\}, \{x\}\}$$

except $\varnothing$ and U; for continuous values, it introduces the idea of "split point". Suppose one attribute of a sample has n continuous values, it then has $n - 1$ splitting points where each of them is the average of two consecutive values $(a[i] + a[i + 1])/2$. Partitions of each attribute are sorted by the amount of impurities that they can reduce. The reduction of impurities could use the most popular method of impurity metric which is: Gini index. If we use k ($k = 1, 2, 3, \cdots, C$) to represent the class, where C is the dependent variable number of the category set. Thus, the Gini impurity of a Node A could be defined as [6]:

$$Gini(A) = 1 - \sum_{k=1}^{C} p_k^2$$

Where p_k denotes the probability of observation points which belong to class k. When $Gini(A) = 0$, all samples belong to the same class. When $Gini(A)$ is the maximum, which is $\frac{(C-1)C}{2}$, all classes occur with the same probability in nodes.

The second key idea in the CART process is to prune the trees of the training set with independent validation data sets. Analyzing the recursive tree construction of classification and regression tree, it is easy to find that there exists a data over-fitting problem [1]. In the construction of decision tree, many branches reflect the abnormality in training data due to the noises or outliers inside. Using such decision tree to classify the data with unknown categories, the accuracy of classification is not high. So it is essential to detect and subtract these branches. Generally, tree pruning method uses statistical metrics, subtract the least reliable branches, which results in faster classification and improves the ability to separate correctly from the training data. The CART algorithm often adopts the post-pruning method, which is implemented by pruning the branches in a fully grown tree. By deleting the branch of the node to cut tree nodes, the bottom non-pruned node becomes a leaf.

The following part of codes shows how we called CART algorithm in our experiment.

```
# Import Library
from sklearn import tree

def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")

    # Train the model using the training sets and check
        score
    model_acc(data, label, dt_model)
```

3.2.2 *Advantage and Disadvantage*

Decision Tree has advantages as follow [4]:

(1) Decision trees are easy to understand and implement, and people have the ability to understand what the decision tree means by explaining it.
(2) Data preparation is often simple or unnecessary for decision trees, and other techniques often require first generalizing data, such as removing redundant or blank attributes.
(3) Feasible and effective results for large data sources in a relatively short period of time.
(4) Not sensitive to missing values
(5) Can handle irrelevant feature data
(6) High efficiency. Decision tree only needs to build once. The maximum number of calculations for each prediction does not exceed the depth of the decision tree

Decision Tree also has disadvantages as follow [4]:

(1) Hard to predict features with continues value
(2) Need to do a lot of data reprocessing work for time-series data
(3) When the category is too large, the error rate may increase.
(4) It does not look good when dealing with data that has a strong correlation between each feature.

3.2.3 *Result*

[Table 1 about here.]

From table 1, we can find that the Decision Tree algorithm has a highest accuracy 0.8378 when we using Clean data. That's because the Clean data contains all 784 features in the data set. It has the minimum information loss among all three data set. Clean data also have the longest running time, which is 20 seconds.

PCA Clean Data have the second highest accuracy with the lowest running time. By using the PCA to reduce the dimension of the clean data, the running time reduced a lot. The accuracy only decreases by 0.01, which shows that the process of PCA did not lose a lot of information.

When we use decision tree algorithm, PCA data have the lowest accuracy. That's may because the raw data have may noise and redundant information. After we remove this information from our data pre-processing step, our accuracy increased.

3.3 Naive Bayes

3.3.1 Introduction

Naive Bayes algorithm is a classification technique based on Baye's Theorem with an assumption of independence among predictors [15]. That is to say, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, we may guess a fruit is an orange if it is yellow, round and about 3 inches in diameter. Even if these features depend on each other, all properties independently contribute to the probability that this fruit is an orange, which explain the term 'Naive' [15].

The Baye's Theorem is particularly useful and not complicated. It solves many problems encountered in our life. The purpose of this theorem is that given a conditional probability of a certain condition, obtain the probability of exchanging two conditions. That is, to get $P(B|A)$ while given $P(A|B)$. $P(A|B)$ is the posterior probability which is also the conditional probability (likelihood) and $P(A)$ or $P(B)$ is called a prior probability. We use the following equation to express this theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The idea of Baye's Theorem is very simple and directly: For the given item which need to be classified, compute the probability of each category under this item. We consider this item belongs to the category with the largest value. The work process of Naive Bayes classification is as follows [20]:

(1) Let D be the set of training tuples associated with their class labels. Each tuple is represented by an n-dimensional attribute vector $X = x_1, x_2, \cdots, x_n$;
(2) Suppose there are m classes $C1, C2, ...Cm$. For the given tuple X, the classification algorithm will predict that X belongs to the class with the highest posterior probability. That is, Naive Bayes classification predicts that X belongs to class C_i if and only if $P(C_i|X) > P(C_j|X), 1 \leq j \leq m, j \neq i$. Thus, the class C_1 with the largest$P(C_i|X)$ is called the maximum posterior probability according to the Baye's Theorem: $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$;
(3) Since $P(X)$ is a constant for all classes, we only require the maximum of $P(C_i|X)P(C_i)$. If the prior probability of a class is unknown, then generally assume these classes are equiprobable (i.e. $P(C_1) = P(C_2) = \cdots = P(C_m)$) and maximize $P(C_i|X)$ based on this assumption. Otherwise, maximize $P(C_i|X)P(C_i)$;
(4) Given a data set with multiple attributes, the computational cost of $P(C_i|X)$ is very large. In order to reduce this cost, we could make the naive assumption about conditional independent of the class. For the label of a given tuple class, assuming the attribute values are conditionally independent. Therefore, we have

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

To examine whether the attribute is classified or continuous value, we need to consider the following two cases:

(a) If A_k is a classified attribute, then $P(x_k|C_i)$ is the number of tuples of class C_i whose value is x_k for attribute A_k in D divided by the number of tuples of class C_i in D ($|C_i, D$);
(b) If A_k is a continuous value attribute, then assume the attribute obeys a Gaussian distribute with the mean η and standard deviation σ, as defined by:

$$g(x, \eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\eta)^2}{2\sigma^2}}$$

Thus, $P(x_k|C_i) = g(x_k, \eta_{C_i}, \sigma_{C_i})$.

(5) To predict the label of class X, calculate $P(C_i|X)P(C_i)$ for each class C_i.

The whole Naive Bayes classification could be divided into three stages:

(1) Preparation stage. The task of this stage is to make the necessary preparation for the Naive Bayes classification. The main work is to determine the characteristic attributes according to the specific situations and make the appropriate partition for each characteristic attribute, and then manually classified some of the items to constitute a training sample set. The input of this stage is all data that need to be classified and the output is the characteristic attribute and the training sample.
(2) Classifier training stage, the task of this stage is to generate a classifier. The main work is to compute the occurrence frequency of each class in training sample and the conditional probability of all partitions in each category, and then record the results. The input is characteristic attributes and a training sample, the output is a classifier. This stage could be completed automatically by a program.
(3) Application stage. The task of this stage is to classify items using classifier. The input is classifier and items, and the output is the mapping between items and categories. This stage could also be completed by a program.

The following part of codes shows how we called Naive Bayes algorithm in our experiment.

```
# Import Library
from sklearn.naive_bayes import GaussianNB

def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")

    # Train the model using the training sets and check
        score
    model_acc(data, label, nb_model)
```

3.3.2 Advantage and Disadvantage

Naive Bayes has advantages as follow [10]:

(1) Naive Bayesian model originated in classical mathematical theory, which is stable.
(2) Have a good performance on small-scale data,
(3) Can handle multi-category tasks.

(4) For incremental training, especially when the amount of data exceeds memory, we can use batch training to save training time.

Naive Bayes also has disadvantages as follow [10]:

(1) In theory, the naive Bayes model has the smallest error rate compared to other classification methods. However, this is not always the case. This is because the naive Bayesian model assumes that the features are independent of each other. This assumption often does not hold in practice. When the number of attributes is large or the correlation between attributes is large, the error rate will be huge.
(2) Need to know the prior probability, and the probability of prior probability depends on the assumption. There are many kinds of hypothetical models, so the prediction results will be poor at some time due to the choice of hypothetical model.
(3) Because we determine the posterior probability by priority and data to determine the classification, there is a certain error rate in the classification decision.
(4) Sensitive to the type of raw data.

3.3.3 *Result*

[Table 2 about here.]

From table 2, we can find that Clean Data have a really low accuracy with the highest time spent. That's because the raw data set did not match the assumption of Naive Bayes. The features are not conditionally independent of each other. The pixels are continues. For example, if pixel1 and pixel3 are both greater than 0, pixel2 will have a more probability to have a value greater than 0.

After we use the dimension reduction technique to reduce the dimension of the data, each component of the data becomes a linear combination of the original data. The new data fits the assumption of Naive Bayes more. Therefore, the PCA Data and PCA Clean Data have a much better performance than Clean Data. They also have the lowest running time compare to any other algorithms.

The PCA Clean Data have the highest accuracy of 0.8710 which higher than the PCA Data. That's may because of the noise and redundant in the original data.

3.4 Logistic Regression

3.4.1 *Introduction*

Logistic regression is a static regression model with a category of the dependent variable. It uses a binary logistic model to estimate binary response probability on predictor variables. In this case, we can know which specific factor makes influence in the presence of risk increasing odds when getting outcomes. We use logistic regression to find the best fitting model to conclude the relationship between variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of the presence of the characteristic of interest [18]:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

p is is the probability of the presence of the characteristic of interest and odds is logical transformation.

$$adds = \frac{p}{1-p} = \frac{p(\text{presence of characteristic})}{p(\text{absence of characteristic})}$$

$$logit(p) = \ln \frac{p}{1-p}$$

There are four ways to input independent variables into the model:

(1) Enter: enter all variables at the same time
(2) Forward: enter essential variables one by one
(3) Backward: enter all variables first and delete non-essential variables one by one
(4) Stepwise: enter essential variables one by one and check the importance of each variable, delete non-essential ones.

It still has other options:

(1) Remove variable. Variables would be removed from the model if its significant level is greater than P-value.
(2) Classification table cutoff value: a value between 0 and 1 which will be used as a cutoff value for a classification table. The classification table is a method to evaluate the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at the selected cut-off value) are cross-classified [18].
(3) Categorical: Identify variables in the category.

The following part of codes shows how we called Logistic Regression algorithm in our experiment.

```
# Import Library
from sklearn.linear_model import LogisticRegression

def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")

    # Train the model using the training sets and check
        score
    model_acc(data, label, lr_model)
```

3.4.2 *Advantage and Disadvantage*

Logistic Regression has advantages as follow [11]:

(1) Very simple to implement and use, widely used in industrial issues
(2) The amount of computation is very small when classified. Therefore the running time is low and the requirement for the storage space is also low.
(3) The sigmoid score for each sample is easy to observe. The threshold can be easily determined by user.
(4) For logistic regression, multicollinearity is not a problem, it can be solved in conjunction with L2 regularization;

Logistic Regression also has disadvantages as follow [11]:

(1) When the feature space is large, the performance of logistic regression is not very good.
(2) May have the under-fitting problem, the general accuracy is not high.

(3) Can only deal with the binary classification problem (based on this, softmax can be used for multi-classification), and must be linearly separable.
(4) For non-linear features, normalization is required.

3.4.3 Result

[Table 3 about here.]

The result of logistic regression is pretty impressive. This is a 10-categorical classification problem, and logistic regression did a good job on this task.

When we get this result, we are thinking if we having an over-fitting result. Therefore, we add a regularization parameter to penalize the features. We use l2 regularization as our parameter when we create our logistic classifier. We also use cross-validation skill to increase our sample size. The results show that the accuracy is still around 90%. Therefore, we are not having an over-fitting problem.

The running time of logistic regression is relatively high. For Clean Data, it received the accuracy of 0.9064 with 218 seconds. PCA Data and PCA Clean Data have a lower accuracy with a much lower time spend. Also, we noticed that the PCA Data accuracy is a little bit higher than the PCA Clean Data. That's may because the clean data make some of the information loss in the raw data.

3.5 Random Forest

3.5.1 Introduction

Random forest uses a random way to build a forest within many decision trees. There is no correlation between each tree in a random forest [21]. After getting the forest, when a new input sample comes in, each decision tree required to make a judgment separately in order to see which class the sample belongs to (for the classification algorithm), and predict the sample for the category which has most selected.

Random forest is mainly used for regression and classification. It is somewhat similar to the bagging which utilizes decision trees as a basic classifier. Bagging could generate a decision tree after replay a sample in each bootstrap and do not make more intervention while generating these trees. Random forest is also sampling with bootstrap, but the difference is that when constructing each tree, every node variable is generated only in a small number of randomly selected variables. Therefore, not only the samples are random, but also the generation of each node's features. Since the combination classifier is more effective than the single classifierThrandom forest could classify the data and give the importance evaluation of each variable.

The basic principle of random forest is to get a new training sample set by selecting k samples from the original training sample set N, and then make up a random forest according to k classification trees. The classification result of the new data depends on the score of the tree votes [14]. In essence, it is an improvement on the decision tree algorithm: it combines multiple decision trees, each tree established depends on an independently sample and has the same distribution. The classification error relies on each the classification ability of a tree and the correlation between them. Feature selection uses a random method to split each node, and then compare the error generated in different situations. The inherent estimation error, classification ability and relevance determine the number of features [14].

Since there are many decision trees in the forest, once a new input sample comes in, each decision tree make a decision to check what the class the sample belongs to, and which one is chosen most to the prediction. There are two selection metrics for decision trees to split attributes [19]:

(1) Information gain
  (a) $I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$, where S is the data set, m is the number of categories, $p_i \approx \frac{|S_i|}{|S|}$ is the probability for any sample belongs to C_i, C_i is a class label and s_i is the number of samples on C_i;
  (b) The smaller $I(s_1, s_2, \cdots, s_m)$, the more ordered of the sample and the better the classification effect;
  (c) Entropy of the subsets partitioned by attribute A: A has V different values, S is partitioned by A into V subsets $s_1, s_2, \cdots, s_v$, where s_{ij} is the number of samples of C_i in subset s_j. Then, we have
$$E(A) = \sum_{j=1}^{V} \frac{(s_{1j} + \cdots + s_{mj})}{s * I(s_{1j}, \cdots, s_{mj})}$$
  (d) $G = I(s_1, s_2, \cdots, s_m)E(A)$;
  (e) Select the attribute with the maximum information gain as the split attribute.
(2) Gini index
  (a) Set S contains N categories of records, then its Gini index is the frequency of the occurrence of p_j;
  (b) If set S is partitioned into m parts $s_1, s_2, \cdots, s_m$, this segmentation is the Gini split;
  (c) Select the attribute with the smallest Gini split as a split attribute.

In order to implement random forest, we should follow these steps:

(1) The input original training set is N, use bootstrap to extract k samples randomly and build k decision trees;
(2) Suppose there are m_A variables, then randomly extract m_T variables from each node of each tree to find one of the variables with the highest classification ability in m_T variables. The threshold of the variable classification is determined by checking each classification point;
(3) Maximize the growth of each tree without any pruning;
(4) Constitute the random forest with these decision trees. Use random forest to determine and classify the new data, and the results are based on votes amount of the tree classifier.

The following part of codes shows how we called Random Forest algorithm in our experiment.

```
# Import Library
from sklearn.ensemble import RandomForestClassifier

def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")

    # Train the model using the training sets and check
        score
    model_acc(data, label, rf_model)
```

3.5.2 Advantage and Disadvantage

Random Forest has advantages as follow [2]:

(1) It can handle very high-dimensional data, and do not have to do feature selection, feature subset is randomly selected
(2) It can provide which feature is more important after training.
(3) When creating a random forest, the use of generalization error is an unbiased estimation, which shows that this model has a high generalization ability.
(4) Easy to make a parallel method, training tree and tree are independent of each other.
(5) In the training process, the algorithm is able to detect the interaction between the features.
(6) For unbalanced data sets, it can balance the model automatically.
(7) If a large part of the features is lost, the model can still maintain the accuracy.

Random Forest also has disadvantages as follow [2]:

(1) There may be many similar decision trees that mask the real results.
(2) Small data or low dimensional data may not produce the best classification.
(3) Much slower than single decision tree algorithm.
(4) Random forests can be over-fitting on some noisy classifications or regression problems
(5) For feature with different value range, the more value-separated features will have a greater impact on random forests

3.5.3 Result

[Table 4 about here.]

From table 4, we can find that Clean Data performed perfectly in this case. It takes the shortest time and reached a 0.9647 accuracy.

The result shows an interesting phenomenon: Clean Data cost less time than PCA Data and PCA Clean Data. In order to explain this phenomenon, we have to check what parameter we choose when we build our random forest classifier. From sklearn API document, we can find that the first default parameter is the number of trees in the forest. For all the data, we set the number of trees to the default number, which is 10. However, in Clean Data, many features are correlated to each other, which means that there may many similar decision trees. For PCA Data and PCA Clean Data, most of the features are independent of each other. Therefore, the running time for Clean Data is higher than PCA Data and PCA clean Data.

Also, we know that when there are similar decision trees in the random forest, the real results may be masked. Therefore, although the Clean Data have a really high accuracy, it may still not as good as the PCA Data and PCA Clean Data result. When we running the classifier on an untested data set, the classifier made by PCA Clean Data may have the best performance among the three.

3.6 Support Vector Machine

3.6.1 Introduction

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [22]. It is mostly used in classification. People can plot each data as a point in an n-dimensional space and give each feature a value. Finding the hyperplane which can differentiate two classes very well can complete classification. As for hyperplane, we must know the notation used to define a hyperplane [12]:

$$f(x) = \beta_0 + \beta^T x$$

β is weight and β_0 is bias.The optimal hyperplane can be represented in an infinite number of different ways by scaling of β and β_0. The one we choose is [12]:

$$|\beta_0 + \beta^T x| = 1$$

x is the training sample who is the most closest to hyperplane. It is known as canonical hyperplane. Distance between point and hyperplane is [12]:

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{||\beta||}$$

$$\text{distance}_{\text{support vector}} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

$$M = 2 * \text{distance}_{\text{support vector}} == \frac{2}{||\beta||}$$

$$minL(\beta) = \frac{1}{2}||\beta||^2 \text{ subject to } y_i(\beta^T + \beta_0) >= 1, \forall i$$

In Python, scikit-learn is a widely used library for implementing machine learning algorithms, SVM is also available in the scikit-learn library and follows the same structure (Import library, object creation, fitting model and prediction). Let's look at the below code [16]:

```
# Import Library
from sklearn import svm

# Assumed you have, X (predictor) and Y (target) for
     training data set and x_test(predictor) of test_data
     set

# Create SVM classification object
model = svm.svc(kernel='rbf', c=10)

# there is various option associated with it, like changing
     kernel, gamma and C value
# Train the model using the training sets and check score
model.fit(X, y)
model.score(X, y)

# Predict Output
predicted= model.predict(x_test)
```

The e1071 package in R is used to create Support Vector Machines with ease. It has helper functions as well as code for the Naive Bayes Classifier. The creation of a support vector machine in R and Python follow similar approaches, let's take a look now at the following code [16]:

```
# Import Library
require(e1071) #Contains the SVM

Train <- read.csv(file.choose())
Test <- read.csv(file.choose())

# there are various options associated with SVM training;
     like changing kernel, gamma and C value.

# create model
model <-
     svm(Target~Predictor1+Predictor2+Predictor3,data=Train,
kernel='linear',gamma=0.2,cost=100)

# Predict Output
preds <- predict(model,Test)
table(preds)
```

3.6.2 *Advantage and Disadvantage*

Support vector machine has advantages as follow:

(1) More efficient in high dimensional space.
(2) Effective when the number of samples is smaller than the number of dimensions.
(3) Can memorize efficiently by using a subset of training sample in decision function.
(4) Flexible by changing Kernel functions for different customers.

And it also has disadvantages as follow:

(1) It would over-fitting in choosing Kernel functions when the number of samples is much smaller than the number of features.
(2) Must pay more attention to regularization term.
(3) It can only get probability by an expensive five-fold cross-validation instead of calculating directly.

3.6.3 *Result*

[Table 5 about here.]

By using SVM to build our classifier, we received a really great accuracy score. The PCA Data received a 0.9814 accuracy of 127 seconds. The running complexity of SVM is $O(N^3 + LN^2 + d * L * N)$, which N is the number of support vector choose, L is the number of samples and d is the number of features of the data set. Therefore, SVM algorithm will run really slow on the large data set. Therefore, when we use the Clean Data, which include more than 42000 samples and 784 features, it takes 1029 seconds to finish the job. We also try to use SVM direct on our raw data. It takes forever to get a result.

SVM can get much better results than other algorithms in the small sample training set. SVM has become one of the most commonly used and effective classifiers. By using the concept of margin, a structured description of the data distribution is obtained, thereby reducing the need for data size and data distribution.

SVM model has three very important parameters kernel, C and gamma[8].

(1) Kernel: string, optional. This parameter specifies the kernel type to be used in the algorithm. There are many different kernels that can be used in SVM. For example, linear, polynomial, sigmoid, Radial basis function (RBF) and precomputed. In this project, choose to use RBF. Because:
   (a) The RBF kernel function can map a sample to a higher dimensional space, and the linear kernel function is a special case of RBF. That is to say, if RBF is considered, then it is unnecessary to consider the linear kernel function.
   (b) Compared with polynomial kernel function, RBF needs to determine fewer parameters, the number of kernel function parameters directly affect the complexity of the function. In addition, when the order of the polynomial is relatively high, the elemental values of the kernel matrix will tend to positive infinity or negative infinity, while the RBF will reduce the numerical calculation difficulties.
   (c) RBF and sigmoid have similar performance for some parameters.
(2) C is the penalty coefficient, which shows the tolerance of the bias. If your C is small, it will give you a great distance, but as a trade-off, we have to ignore some misclassified samples; on the other hand, if you have a large C, you will try to correctly classify all the samples, but the price is the margin space will be small. In our example, we choose c equals to 10, which is a relatively large c value, which brings us a more accurate classifier.
(3) Gamma defines how much influence a single training example has. It determines the distribution of the data after mapping to a new feature space. The larger the gamma is, the less the support vector it will be. The smaller the gamma value is, the more the support vector it will be. The number of support vectors affects the speed of training and prediction. Also, if we set gamma large, it will have the over-fitting problem. Therefore, in this project, we decided to use the default gamma value, which is

$$gamma = \frac{1}{\text{number of features}}$$

In this task, SVM have a really great performance, the running time is also acceptable.

4 CONCLUSION

[Table 6 about here.]

From table 6, we can easily find that when we use SVM classifier on PCA Data, we will receive the highest accuracy among all 5 different algorithms. The highest accuracy we reached for this project is 0.9813, which shows that our classifier predicts 98.13% of the sample correct by using our SVM classifier. The time of training the model takes 127 seconds. The time spent is acceptable. The accuracy of SVM on PCA Clean Data has the second highest accuracy, which is 0.9785. The difference between first and second highest accuracy is about 0.0028, which is really small. However, the time spent saved 41.1%. Therefore, SVM on PCA Clean Data is also a reasonable choice for the Digit Recognition task.

Random Forest can be explained as a combination of many decision trees. Decision tree can be explained as a special case of Random Forest, which set the number of trees in the Random Forest to 1. Therefore, Random Forest has a much better performance than decision tree in all three data set. As a trade-off, the time spent for Random Forest is much higher than Decision Tree.

Compare to other four Classifiers, Naive Bayes has the fastest training speed. For PCA Data and PCA Clean Data, Naive Bayes Classifier takes less than one second to train the classifier. And for Clean data, which contains all 784 features, it takes only 6 seconds to train the classifier. The reason why Naive Bayes is fast is that:

(1) The algorithm does not need to iterate to get the result. The running time is approximately linear.
(2) It makes an assumption of independence between its features, so that parameter estimates can be calculated independently and thus possibly very quickly.
(3) The prior probability values do not change. Therefore, the prior probability can be calculated and store in memory in the first place.

However, we have to be very careful about the assumption made by Naive Bayes, or we will get a very low accuracy.

Logistic Regression received an average performance among the 5 algorithms. It achieves a 0.8891 accuracy in 27 seconds on PCA data. However, when we using logistic regression, we have to pay a lot of attention to over-fitting problem. We should use regularization and cross-validation to reduce the probability of over-fitting problems.

To conclude, we decide to use SVM classifier for Digit Recognition Task. We should definitely use feature extraction on the data because of the running time and over-fitting problem. The Binary Data cleaning method is optional. If we want to have higher accuracy, we should not use Binary Data cleaning. As a trade-off, if we want to have faster training speed, we should use Binary Data cleaning.

5 FUTURE IMPROVEMENT

Our analysis is far from perfect. There are several points that we want to point our as discussion and also opportunities for future improvement.

(1) We can try several more classification algorithms. For example, K^{th} Nearest Neighbour (KNN) and Neural Network. We can use some more complex algorithms too, such as Convolution Neural Network (CNN).
(2) We can focus more on tune parameter. For example, we can use the Grid Search on SVM to get a better parameter combination.
(3) We can choose a different Data Cleaning Method. For example, we can set a threshold on data. Any value greater than 50 will be set to 1.
(4) We can choose a different Feature Extraction or Feature Selection method. For example, LDA. Unlike PCA, LDA is an unsupervised dimension reduction method.

ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski and all TAs for providing the resource, tutorials and other related materials to write this paper.

REFERENCES

[1] Jason Brownlee. 2016. Classification And Regression Trees for Machine Learning. (April 2016). https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/
[2] Daniel S. Chapman, Aletta Bonn, William E. Kunin, and Stephen J. Cornell. 2009. Random Forest characterization of upland vegetation and management burning from aerial imagery. *Journal of Biogeography* 37, 1 (2009), 37–46. https://doi.org/10.1111/j.1365-2699.2009.02186.x
[3] Kazuhiro Hotta. 2008. Non-linear feature extraction by linear PCA using local kernel. *2008 19th International Conference on Pattern Recognition* (2008). https://doi.org/10.1109/icpr.2008.4761721
[4] Hemant Ishwaran and J. Sunil Rao. 2009. Decision Tree: Introduction. *Encyclopedia of Medical Decision Making* (2009). https://doi.org/10.4135/9781412971980.n97
[5] Kaggle. 2015. Data Discription. (2015). https://www.kaggle.com/c/digit-recognizer/data
[6] Scikit Learn. 2007. Decision Trees. (2007). http://scikit-learn.org/stable/modules/tree.html
[7] Scikit Learn. 2007. sklearn model selection cross val score. (2007). http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html
[8] Scikit Learn. 2007. sklearn svm SVC. (2007). http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[9] Sheng-Jie Liang, Zhi-Hua Zhang, and Li-Lin Cui. 2010. Feature extraction method Based PCA and KICA. *2010 Second International Conference on Computational Intelligence and Natural Computing* (2010). https://doi.org/10.1109/cinc.2010.5643821
[10] J. Luengo and Rafael Rumi. 2015. Naive Bayes Classifier with Mixtures of Polynomials. *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (2015). https://doi.org/10.5220/0005166000140024
[11] Scott Menard. 2010. Introduction: Linear Regression and Logistic Regression. *Logistic Regression: From Introductory to Advanced Concepts and Applications* (2010), 1–18. https://doi.org/10.4135/9781483348964.n1
[12] OpenCV. 2017. Introduction to Support Vector Machines. (December 2017). https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
[13] OpenML. 2013. 10-fold Crossvalidation. (2013). https://www.openml.org/a/estimation-procedures/1
[14] Savan Patel. 2017. Chapter 5: Random Forest Classifier. (May 2017). https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1
[15] Sunil Ray. 2015. 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R). (September 2015). https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[16] Sunil Ray. 2015. Understanding Support Vector Machine algorithm from examples (along with code). (October 2015). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
[17] Dr. Saed Sayad. 2010. Decision Tree - Classification. (2010). http://www.saedsayad.com/decision_tree.htm
[18] MedCalc Software. 2017. Logistic regression. (February 2017). https://www.medcalc.org/manual/logistic_regression.php
[19] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Addison Wesley.
[20] Wikipedia. 2017. Naive Bayes classifier. (December 2017). https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[21] Wikipedia. 2017. Random forest. (November 2017). https://en.wikipedia.org/wiki/Random_forest
[22] Wikipedia. 2017. Support vector machine. (December 2017). https://en.wikipedia.org/wiki/Support_vector_machine
[23] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar. 2005. Enhancing Data Analysis with Noise Removal. (2005). https://doi.org/10.21236/ada439494

A CODE ATTACHMENT

```
##Author: Yuchen Liu HID213, Wenxuan Han HID209, Junjie Lu
    HID214
##Data: 2017.12.01
```

```python
##Reference:
     http://blog.csdn.net/tinkle181129/article/details/55261251

from datetime import datetime
import matplotlib.pyplot as plt
import pandas as pd
from numpy import *
from sklearn import svm
from sklearn import tree
from sklearn.cross_validation import cross_val_score
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB


# 1. read data from csv
def read_data():
    data_set = pd.read_csv("train.csv")
    data = data_set.values[0:, 1:]
    label = data_set.values[0:, 0]
    print("Data load completed.")
    return data, label


# plot 70 samples
def show_pic(data):
    print(shape(data))
    plt.figure(figsize=(7, 7))
    for digit_num in range(0, 70):
        plt.subplot(7, 10, digit_num + 1)
        grid_data = data[digit_num].reshape(28, 28)
        plt.imshow(grid_data, interpolation="none",
             cmap="afmhot")
        plt.xticks([])
        plt.yticks([])
    plt.tight_layout()
    plt.savefig("data_samples.png")


# 2. Data Cleaning
# The data is from 0-255 for each cell.
# Normalize data by set all value > 0 to 1
def data_clean(data):
    m, n = shape(data)
    new_data = zeros((m, n))
    for i in range(m):
        for j in range(n):
            if data[i, j] > 0:
                new_data[i, j] = 1
            else:
                new_data[i, j] = 0
    print("Data clean completed.")
    return new_data


# 3. Feature Selection by PCA
def feature_selection(data):
    # First, use explained_variance to get recommended
         number of component
    pca = PCA()
    # pca_parameter = pca.fit(data)
    pca.fit(data)
    ev = pca.explained_variance_
    ev_ratio = []
    for i in range(len(ev)):
        ev_ratio.append(ev[i] / ev[0])

    # select number of component which have a higher ratio
    # than 0.05 with the first components
    n = 0
    for i in range(len(ev_ratio)):
        if ev_ratio[i] < 0.05:
            n = i
            # print(n)
            break

    # Then, PCA the model by the number of components
    # pca = PCA(n_components=n, whiten=True)
    pca = PCA(n_components=n, whiten=True)
    print("Feature selection completed.")
    return pca.fit_transform(data)


# 4. Model Selection
def model_acc(data, label, model):
    start = datetime.now()
    acc = cross_val_score(model, data, label,
         scoring="accuracy").mean()
    end = datetime.now()
    time_use = (end - start).seconds
    print("Time use: ", time_use)
    print("Accuracy by cross validation: ", acc)


def dt_classifier(data, label, data_type):
    dt_model = tree.DecisionTreeRegressor()
    dt_model.fit(data, label)
    print("Test " + data_type + " using DT: ")
    model_acc(data, label, dt_model)


def nb_classifier(data, label, data_type):
    nb_model = GaussianNB()
    nb_model.fit(data, label)
    print("Test " + data_type + " using NB: ")
    model_acc(data, label, nb_model)


def lr_classifier(data, label, data_type):
    lr_model = LogisticRegression()
    lr_model.fit(data, label)
    print("Test " + data_type + " using LR: ")
    model_acc(data, label, lr_model)


def rf_classifier(data, label, flag):
    rf_model = RandomForestClassifier(n_estimators=100)
    rf_model.fit(data, label)
    print("Test " + flag + " using RF: ")
    model_acc(data, label, rf_model)


def svm_classifier(data, label, flag):
```

```
    svm_model = svm.SVC(kernel="rbf", C=10)
    svm_model.fit(data, label)
    # svc_clf = NuSVC(nu=0.1, kernel='rbf', verbose=True)
    print("Test " + flag + " using SVM: ")
    model_acc(data, label, svm_model)


def main():
    data, label = read_data()
    # show_pic(data)
    clean_data = data_clean(data)

    test_type = 3
    for i in range(1, 3):
        print("In %d test" % i)

        if test_type == 0:
            input_data = data
            str = "raw data"
        elif test_type == 1:
            input_data = clean_data
            str = "clean data"
        elif test_type == 2:
            input_data = feature_selection(data)
            str = "pca data"
        elif test_type == 3:
            input_data = feature_selection(clean_data)
            str = "pca clean data"

        dt_classifier(input_data, label, str)
        nb_classifier(input_data, label, str)
        lr_classifier(input_data, label, str)
        rf_classifier(input_data, label, str)
        svm_classifier(input_data, label, str)


main()
```

List of Figures

1 70 samples of hand-drawn digits in this data set. 14

Figure 1: 70 samples of hand-drawn digits in this data set.

List of Tables

1 Result For Decision Tree 16
2 Result For Navie Bayes 16
3 Result For Logistic Regression 16
4 Result For Random Forest 16
5 Result For Support Vector Machine 17
6 Result For Different Algorithm with Different Data Cleaning & Feature Extraction method 17

	PCA Data	PCA Clean Data	Clean Data
Time	12	9	20
Accuracy	0.8012	0.8234	0.8378

Table 1: Result For Decision Tree

	PCA Data	PCA Clean Data	Clean Data
Time	0	0	20
Accuracy	0.8651	0.8710	0.5397

Table 2: Result For Navie Bayes

	PCA Data	PCA Clean Data	Clean Data
Time	27	21	218
Accuracy	0.8891	0.8862	0.9064

Table 3: Result For Logistic Regression

	PCA Data	PCA Clean Data	Clean Data
Time	126	107	56
Accuracy	0.9483	0.9497	0.9647

Table 4: Result For Random Forest

	PCA Data	PCA Clean Data	Clean Data
Time	127	90	1029
Accuracy	0.9814	0.9785	0.9575

Table 5: Result For Support Vector Machine

	Decision Tree	Naive Bayes	Logistic Regression	Random Forest	Support Vector Machine
PCA Time	12	**0**	27	126	127
PCA Accuracy	0.8012	0.8651	0.8891	0.9483	**0.9813**
PCA Clean Time	9	**0**	21	107	90
PCA Clean Accuracy	0.8234	0.8710	0.8862	0.9497	0.9785
Clean Time	20	6	218	56	1029
Clean Accuracy	0.8378	0.5397	0.9064	0.9647	0.9575

Table 6: Result For Different Algorithm with Different Data Cleaning & Feature Extraction method

```
bibtext report
==========================================

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--entry type for "cart" isn't style-file defined
--line 5 of file report.bib
Warning--entry type for "sklearn.cv" isn't style-file defined
--line 57 of file report.bib
Warning--entry type for "sklearn.dt" isn't style-file defined
--line 68 of file report.bib
Warning--entry type for "svm.form" isn't style-file defined
--line 117 of file report.bib
Warning--entry type for "10f.cv" isn't style-file defined
--line 129 of file report.bib
Warning--entry type for "sp.rfc" isn't style-file defined
--line 139 of file report.bib
Warning--entry type for "nb.steps" isn't style-file defined
--line 150 of file report.bib
Warning--entry type for "svm.code" isn't style-file defined
--line 162 of file report.bib
Warning--entry type for "ss.dt" isn't style-file defined
--line 174 of file report.bib
Warning--entry type for "lr.form" isn't style-file defined
--line 185 of file report.bib
Warning--entry type for "wiki.nb" isn't style-file defined
--line 208 of file report.bib
Warning--entry type for "wiki.rf" isn't style-file defined
--line 219 of file report.bib
Warning--entry type for "wiki.svm" isn't style-file defined
--line 231 of file report.bib
Warning--no number and no volume in PCA
Warning--page numbers missing in both pages and numpages fields in PCA
Warning--no number and no volume in DT
Warning--page numbers missing in both pages and numpages fields in DT
Warning--no number and no volume in feature_extra
Warning--page numbers missing in both pages and numpages fields in feature_extra
Warning--no number and no volume in NB
Warning--page numbers missing in both pages and numpages fields in NB
Warning--no number and no volume in LR
Warning--empty address in intro.dm
Warning--no journal in data_clean
Warning--no number and no volume in data_clean
```

```
Warning--page numbers missing in both pages and numpages fields in data_clean
(There were 26 warnings)

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

report.bib:243:@Article{data_clean,
report.bib:89:@Article{feature_extra,

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


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


latex report
===========================================

        [2017-12-04 22.00.29] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.3s.
../README.yml
  24:81     error    line too long (82 > 80 characters)  (line-length)
  45:81     error    line too long (85 > 80 characters)  (line-length)
  47:79     error    trailing spaces  (trailing-spaces)
  48:81     error    line too long (82 > 80 characters)  (line-length)
  48:82     error    trailing spaces  (trailing-spaces)
  49:79     error    trailing spaces  (trailing-spaces)
  50:81     error    line too long (81 > 80 characters)  (line-length)
  50:81     error    trailing spaces  (trailing-spaces)
  51:81     error    line too long (89 > 80 characters)  (line-length)
  51:89     error    trailing spaces  (trailing-spaces)
  52:81     error    line too long (81 > 80 characters)  (line-length)
  52:81     error    trailing spaces  (trailing-spaces)
  53:81     error    line too long (81 > 80 characters)  (line-length)
  53:81     error    trailing spaces  (trailing-spaces)
  54:81     error    line too long (86 > 80 characters)  (line-length)
  54:86     error    trailing spaces  (trailing-spaces)
```

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


name: Han, Wenxuan
hid:  209
paper1: Oct 29 2017 100%
paper2: 100%
project: Dec 04 17 100%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

17
wc 209 project 17 8949 report.tex
wc 209 project 17 9030 report.pdf
wc 209 project 17 719 report.bib

find "
--------------------------------------------------------------------------------

152: print("Data clean completed.")

243: acc = cross_val_score(model, data, label, cv=5,
     scoring="accuracy").mean()

247: print("Time use: ", time_use)

248: print("Accuracy by cross validation: ", acc)

295: print("Test " + data_type + " using DT: ")

388: print("Test " + data_type + " using NB: ")

469: print("Test " + data_type + " using LR: ")

558: print("Test " + flag + " using RF: ")

794: data_set = pd.read_csv("train.csv")
```

```
797: print("Data load completed.")

808: plt.imshow(grid_data, interpolation="none", cmap="afmhot")

812: plt.savefig("data_samples.png")

827: print("Data clean completed.")

854: print("Feature selection completed.")

861: acc = cross_val_score(model, data, label,
     scoring="accuracy").mean()

864: print("Time use: ", time_use)

865: print("Accuracy by cross validation: ", acc)

871: print("Test " + data_type + " using DT: ")

878: print("Test " + data_type + " using NB: ")

885: print("Test " + data_type + " using LR: ")

892: print("Test " + flag + " using RF: ")

897: svm_model = svm.SVC(kernel="rbf", C=10)

900: print("Test " + flag + " using SVM: ")

911: print("In %d test" % i)

915: str = "raw data"

918: str = "clean data"

921: str = "pca data"

924: str = "pca clean data"

passed: False

find footnote
--------------------------------------------------------------------------------

passed: True
```

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

4: \input{format/i523}

passed: True

find input{format/final}
--------------------------------------------------------------------------------

passed: False

floats
--------------------------------------------------------------------------------

108: \begin{figure}
109: \includegraphics[width=0.35\columnwidth]{images/data_samples}

figures 1
tables 0
includegraphics 1
labels 0
refs 0
floats 1

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

Label/ref check
106: In train.csv data file, it contains $42000$ gray-scale images of
     hand-drawn digits, from zero through nine. Each image is a $28
     \times 28$ pixels matrix with a total of 784 pixels
     \cite{kaggle}. Each pixel has a single pixel-value which is an
     integer from 0 to 255 associated with it, indicating the
     lightness or darkness of that pixel (higher numbers meaning
     lighter). In this experiment, we have plotted the graph in order
     to see the appearance of these digits easily. Figure 1 shows the
     first 70 samples.
332: From table 1, we can find that the Decision Tree algorithm has a
     highest accuracy 0.8378 when we using Clean data. That's because
     the Clean data contains all 784 features in the data set. It has
     the minimum information loss among all three data set. Clean data
     also have the longest running time, which is 20 seconds.
passed: False -> labels or refs used wrong
```

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



WARNING: figure and above may be used improperly

127: As we mentioned above, it can be seen from both the figure and
     the pixel-value that the value varies from 0 to 255, which means
     each feature is a continuous value. Thus, it is possible that
     such continuous values might affect our later feature selection.
     Our observation shows that the values are not very high at the
     boundaries of 0 and $>0$. So here exist three ways to handle it
     \cite{data_clean}:

WARNING: code and below may be used improperly

626: In Python, scikit-learn is a widely used library for implementing
     machine learning algorithms, SVM is also available in the scikit-
     learn library and follows the same structure (Import library,
     object creation, fitting model and prediction). Let's look at the
     below code \cite{svm.code}:

WARNING: algorithm and below may be used improperly

626: In Python, scikit-learn is a widely used library for implementing
     machine learning algorithms, SVM is also available in the scikit-
     learn library and follows the same structure (Import library,
     object creation, fitting model and prediction). Let's look at the
     below code \cite{svm.code}:

bibtex
--------------------------------------------------------------------------------



label errors
```

```
89: feature_extra: do not use underscore in labels:
243: data_clean: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--entry type for "cart" isn't style-file defined
--line 5 of file report.bib
Warning--entry type for "sklearn.cv" isn't style-file defined
--line 57 of file report.bib
Warning--entry type for "sklearn.dt" isn't style-file defined
--line 68 of file report.bib
Warning--entry type for "svm.form" isn't style-file defined
--line 117 of file report.bib
Warning--entry type for "10f.cv" isn't style-file defined
--line 129 of file report.bib
Warning--entry type for "sp.rfc" isn't style-file defined
--line 139 of file report.bib
Warning--entry type for "nb.steps" isn't style-file defined
--line 150 of file report.bib
Warning--entry type for "svm.code" isn't style-file defined
--line 162 of file report.bib
Warning--entry type for "ss.dt" isn't style-file defined
--line 174 of file report.bib
Warning--entry type for "lr.form" isn't style-file defined
--line 185 of file report.bib
Warning--entry type for "wiki.nb" isn't style-file defined
--line 208 of file report.bib
Warning--entry type for "wiki.rf" isn't style-file defined
--line 219 of file report.bib
Warning--entry type for "wiki.svm" isn't style-file defined
--line 231 of file report.bib
Warning--no number and no volume in PCA
Warning--page numbers missing in both pages and numpages fields in PCA
Warning--no number and no volume in DT
Warning--page numbers missing in both pages and numpages fields in DT
Warning--no number and no volume in feature_extra
Warning--page numbers missing in both pages and numpages fields in feature_extra
Warning--no number and no volume in NB
Warning--page numbers missing in both pages and numpages fields in NB
Warning--no number and no volume in LR
Warning--empty address in intro.dm
```

```
Warning--no journal in data_clean
Warning--no number and no volume in data_clean
Warning--page numbers missing in both pages and numpages fields in data_clean
(There were 26 warnings)

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------

non ascii found 65292

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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

Income Prediction based on Machine Learning Techniques

Borga Edionse Usifo
Indiana University
Bloomington, Indiana 47408
busifo@iu.edu

ABSTRACT

This project takes a closer look to some of the most used supervised learning algorithms in machine learning. We start with the description of the each of the algorithms then we move it to analytics and findings by using that particular algorithm in our data-set. We also provide advantages and disadvantages of each supervised machine learning algorithm for future reference. We mainly focus on our prediction of the income level of individuals by looking at their age, gender, education, location, and other features given by our data-set. We will try each algorithm and try to pick the best features from our data-set to have an optimal prediction.

KEYWORDS

i523, HID343, Machine Learning, Income Prediction, Logistic Regression, Ensemble methods

1 INTRODUCTION

In this project, we try to showcase the performance of the machine learning algorithms on data which we gather from UCI machine learning repository [22]. This data used by Kohavi R. and Becker B. for their research in improving the in Naive Bayes Classifier's accuracy [21].

Data consists of 15 variables, and we try to predict the income of the individuals. To do this prediction task, we first started with data preparation because the data we receive from UCI machine learning repository [22] not fully prepared for any machine learning algorithm. Our first task was the clean the data while applying some statistical techniques to get insights from the dataset. We also used data transformation methods like One-Hot-Encoding[45] to apply logarithmic functions for improving the machine learning algorithms performance before training the data.

Machine Learning algorithms that we discuss in this paper are Gaussian Naive Bayes [46], K Nearest Neighbors [29], Ensemble Methods (Boosting) [8], Support Vector Machines [6], Logistic Regression [34], and Decision Trees [49]. We try to show their weakness, advantages, and their time consumption while training each of them in machine learning algorithms section.

After providing a brief introduction of each of the supervised machine learning algorithms, we will discuss our findings for of each of the algorithms by comparing their accuracy score, F-1 score, recall, and lastly time comparison.

2 IMPORTANCE OF BIG DATA ANALYTICS FOR PREDICTIVE CLASSIFICATION

Importance of big data analytics is getting higher every day since the algorithms become more powerful to predict, classify and cluster any given data set. Importance of our case is any company can be used to predict individuals income to refer them goods in their income range or governments can provide additional support for the areas that have lower income range. There can be many possible things that can do with this kind of classification predictions.

3 DATA PREPARATION

We first used the pandas [28] to help to load the data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame [48]. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was "income" attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables (age, education, capital gain, capital loss, hours worked) by using the pandas [27] functions as shown in Table 1.

[Table 1 about here.]

3.1 Data Cleaning

After getting a snapshot from income data frame, we recognized that there is a column which has no meaning. The first task was to remove this entire column from our dataset we used pandas drop function for doing this task. After removing this column, we had more concise dataframe to analyze.

Moreover, removing the column we have encountered some missing values which labeled as "question marks" in data frame. In order to remove this values we first changed all the "question mark" values to "NaN" values by using pandas "replace" function [26]. After replacing all the question marks with "NaN" values, we used pandas missing value dropping function to remove all the "NaN" values from our dataset.

Furthermore, we start investigating the types of the variables, and in our case, we found two types of variable one of them labeled as "int64" which stands for integer values, other one labeled as object type of variable. From our previous example especially in "scikit-learn" it is better to use float object rather than "int64" for training the machine learning algorithms. Because their numerical output most of the time is "float64" object. We transferred all the "int64" objects to "float64" objects. This was the last step of the cleaning process.

Our last process is changing the string values to numerical values on our target data which consist of string values ("$ 50K") for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table 2.

[Table 2 about here.]

Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called

one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well [45]. In other words "we implement binarization of the category to include as a future to train model [45]". As we can see in Table 3 and Table 4.

[Table 3 about here.]

[Table 4 about here.]

4 DATA EXPLORATION

After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than $50K and less than $50K which can be seen in below Table 4.

Description	Count
Total Number of Individuals	30162
Individuals who makes more than $50K	7508
Individuals who makes at or less than $50K	22654

Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table 5. As we can see we have individuals who're age ranging from 17 to 90 years old with a mean of 38.58. If we look at the capital gains and capital losses, we have a standard deviation of 7385 and 402 respectively this is also another indication of skew in these variables.

[Table 5 about here.]

We used scatter matrix plot and applied the correlation function to see if we have any reliable correlation between any of the variables. As we can see from the correlation matrix Table 6 and correlation numbers Figure 1we do not have the high correlation between any variables. Correlation values range between -1 to 1. The correlation value of 1 is an indication of perfect positive correlation and correlation number -1 indicates a negative correlation between variables [15]. Because of lower correlation values, it will be tough to determine the classification by just looking at the correlations; this indicates we have sophisticated algorithms to determine the relationship between variables to classify individuals incomes.

[Table 6 about here.]

[Figure 1 about here.]

Furthermore, we also explore the capital gains, capital losses, and hours per week variables which we used a histogram to plot the data into distribution form so we can see how all these attributes distributed. The reason we do the histogram is we want to see any skewness in our data. As shown in the histogram graphs in Figure 2 and Figure 3 in capital gains and capital loss we have highly skewed data which can cause issues later on in our algorithms. We apply a logarithmic function to do highly skewed data to less skewed [24]. Using logarithmic functions adds more value to data from the interpretable standpoint and "it helps to meet the assumptions of inferential statistics [24]".

[Figure 2 about here.]

[Figure 3 about here.]

Moreover, applying logarithmic function had an impact on distribution. We can see the changes on skew data in Figure 4 after applying logarithmic function.

[Figure 4 about here.]

5 MACHINE LEARNING ALGORITHMS TO CONSIDER

We have multiple algorithms to consider when we are doing the supervised learning. Each algorithm has its benefits and drawbacks. We will consider several supervised machine learning algorithms for our predictions. The application we will use to implement these algorithms will be Python Scikit-Learn library. We will briefly explain each parameter included in these algorithms in Scikit-Learn.

First we'll look at the Scikit-Learn in Python framework we will go through the advantages in Scikit-Learn how we can implement any machine learning in just couple of simple line of codes in Scikit-Learn.

5.1 Why Scikit-Learn?

Scikit-learn developed by David Cournapeau in 2007. The development came from while he was working on summer code project for Google. After recognized and published by INRIA in 2010 project start the get more attention among worldwide. There are more than 30 active contributors and has secured several sponsorships from big technology companies[17]. "It also has a goal of providing common algorithms to Python users through consistent interface[2]". Scikit-Learn consists of several elements to make analytical predictions. These elements are shown below[23]:

Supervised Learning Algorithms: One of the most fundamental reason that Scikit-Learn's popularity comes from highly available supervised learning algorithms. These algorithms vary from regression models to decision trees and many more[23].

Cross Validation: Scikit-Learn includes various techniques to check the accuracy or any statistical measure between training and unseen testing set[23].

Unsupervised Learning Algorithms: Scikit-Learn had also various algorithms to support many unsupervised algorithms some of these include clustering, factor analysis, and neural network analysis[23].

Various example data-sets: Scikit-Learn comes with different data sets included in its package so users can start learning Scikit-Learn without the need of any data-sets[23].

Feature extraction: It has rich feature for extracting images or text from data-sets[23].

Algorithms that we will investigate shown below; we will go more deep analysis on each of these algorithms.

- Gaussian Naive Bayes
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Stochastic Gradient Descent Classifier
- Support Vector Machines
- Decision Trees

5.2 Gaussian Naive Bayes

Naive Bayes bring many beneficial features; it is widely popular among machine learning applications[41]. The popularity of Naive

Bayes comes from being able to handle large projects and data-sets faster than most algorithms[41]. It also can handle complex data-sets with categorical and non-categorical inputs [41]. Naive Bayes based on probabilistic classifier of Bayesian theory. It is also a favorite way of doing text categorization [46].

Term naive comes from it is the method of use probability among categories which assumes of independence among given class of attributes as shown in Figure 5. In other words, if we try to classify individuals from their email communications it will not take the order of words into account. Whereas in the English language we can tell the difference between sentence makes sense or not if we randomly re-order our words in the sentences. So it does not understand the text, it only looks at word frequencies as a way to do the classification. This is why it is called "Naive".

[Figure 5 about here.]

As we state above Naive Bayes derives from Bayesian Theory where the dimensionality of inputs is relatively high. Bayesian Theorem is stated below [16].

$$P(C \mid X) = \frac{P(X \mid C) \times P(C)}{P(X)} \tag{1}$$

Naive Bayes Classifier works as follows [16]:

Advantages of Naive Bayes [16]:

- Faster classification time for training data-set.
- Because of independent classification it improves classification performance.
- Performance is relatively good.

Disadvantages of Naive Bayes[16]:

- Often it requires a large number of data-sets to give adequate results.
- On some occasions which are relative to data-sets, it can give less accuracy.

5.3 Logistic Regression

Logistic Regression widely used for predicting "probability of failure in a given system, product, and process [34]". Logistic Regression also used in natural language analysis, it is an extension of conditional random fields [34]. It works as a classifier which learns the features from the input given and classifies them by multiplying the input value with the weight value [14].

$$P(C \mid X) = \sum_{i=1}^{N} W_i \times f_i \tag{2}$$

Main reason that Logistic Regression differs from Linear Regression is output variable for Logistic Regression is binary whereas output variable in Linear Regression is discrete(continuous) [12].

Advantages of Logistic Regression:

- It does not have any assumptions over distribution of classes [18].
- It is fast to train [18].
- Logistic Regression has fast classifying method of unknown data [18].
- We can easily extend to other regression for multiple classes like multinomial regression [18].

Disadvantages of Logistic Regression:

- One of the disadvantages of linear regression is it is not providing flexibility in some instances. What we mean by the " lack of flexibility is the linear dependency, and linear decision boundary in the instance space is not valid [42]". This disadvantage can be improved changing from Logistic Regression to Choquistic Regression[42].
- Logistic regression can provide poor results when there are more complex relationships in data [9].
- Logistic models also have over-fitting problems which come from a result of sampling bias [31].
- Because of Logistic Regression's predictions comes from the independent variable if the researcher includes wrong independent variables then model's prediction will have no value [31].
- Because it is predictions based on 1 and 0 model will have poor performance when predicting continuous variables [31].

5.4 K-Nearest Neighbors (KNN)

K Nearest neighbor has been primarily studied, and this popularity comes from it has been applied to many applications some of these applications are "spatial databases, pattern recognition, geographic information, image retrieval, computer game, and many other applications [29]". Due to an increase of mobile devices and people tends to use of applications like navigation K-nearest neighbor found itself another widely used area of location-based services due to an ability to found a target location [29].

Intuition behind the K Nearest Neighbor can be described as follows: " for a set P of n objects and a querying point q, return the k objects in P that are closest to q [29]."

Advantages of K Nearest Neighbors:

- K Nearest Neighbor is a basic and simple approach to implement [35].
- K Nearest Neighbor can perform well and efficiently with the large amount of data [43].
- K nearest Neighbor also does effectively well with noisy data sets ("if the inverse square of weighted distance used as the distance [43]"). In other words, it is flexible to feature and distance choices [35].

Disadvantages of K Nearest Neighbors:

- K Nearest Neighbor typically require large dataset to perform well [35].
- Time complexity could be high due to computing distance of each query to all training data points [43]. This time might be improved with some indexing (K-D Tree) [43].
- Determining the value of K can be time-consuming [43].
- It can be unclear to know which type of distance to use, as well as which variability to use to get the optimal results [43].
- Switching the different K values can result in the predicted class labels [30].

Many of these disadvantages are improving with the help of parallel distributed computing. Recent improvements in MapReduce framework allows users to run KNN algorithms in the cluster which had a significant effect on reducing the computation time [19].

Another area of improvements on KNN, is to implement different mapping functions such as kernel KNN, kernel difference weighted KNN, adaptive quasi-conformal kernel nearest neighbor, angular similarity, local linear discriminant analysis, and Dempster-Shafer [10].

5.5 Decision Trees

Decision Tree is another widely used algorithm model for classification and regression. Decision Trees uses a recursive split model where each recursive split is identified by each data point; this is an example of non-parametric hierarchical model [13].

Representation of decision trees is as follows; we sort the instances from root to leaf nodes, this sorting gives insights about the classification of the instance, every outcome descending from the root node corresponds to possible values for that variable [33]. We can classify an instance by starting from the root node and checking the attributes labeled on that node and moving down from that node based on attribute given attribute values [33] as shown in Figure 6.

[Figure 6 about here.]

Advantages of Decision Trees:

- Decision Tree applications are easy to interpret and understand [32]. This ease comes from their schematic representation [32]. Interpretation between alternatives can be expressed with single numerical number which is the expected value (EV) [32].
- Decision Trees can handle noisy or incomplete data-sets [32]. In other words it requires little effort of data preparation because of it is flexibility [7].
- It can handle both nominal and numerical variables [32].
- It can be modified easily whenever the new information is available [32].
-

Disadvantages of Decision Trees:

- Because of it is a use of divide and conquer method they can demonstrate good performance if there are few attributes exists when the attributes level goes into large number decision tree become more complex which will result in poor performance [32].
- Decision Trees are also susceptible to training set which can give a result of over-fitting [32]. In other words, it can believe the training set completely which will give an abysmal performance on testing set.
- ID3 and C4.5 decision tree algorithms require discrete values as input data.

5.6 Stochastic Gradient Descent Classifier (SGD)

Stochastic Gradient Descent recently got became more popular because of it is large-scale learning ability in machine learning problems [11]. It is a useful and straightforward way approach of linear classifiers under convex problems which is Support Vector Machines or Conditional Random Fields [3]. The originality of SGD derives from "Stochastic Approximation" which is a work from Robinson and Manroe [5].

Advantages of Stochastic Gradient Descent:

- One of the advantage of stochastic gradient descent is, it is easy to implement [38].
- Stochastic Gradient Descent is also efficient because of each step only relies on a single derivative which makes the computational cost 1 / n than normal gradient descent [37].

Disadvantages of Stochastic Gradient Descent:

- Stochastic Gradient Descent can be required to have many iterations, and it also requires some hyper-parameters [38].
- Feature scaling is a practice which used in the standardization of range of independent variables [47]. SGD also used this feature scaling technique and it can be sensitive to feature scaling [38].
- Another drawback of Stochastic Gradient Descent is while using GPU they are hard to parallelize or distributing them using computer clusters [25].

5.7 Support Vector Machines

Support Vector Machines is fallen under the classification methods in machine learning [6]. It is also a robust classification method that has been widely found itself an area ranging from pattern recognition to text analysis [6].

Fitting a boundary between data points is the principle of the support vector machines. This boundary divides the data points between classes, and each similar data point puts under the same class classification [6]. After training the support vector machines with training data-set, we only need to check whether the test data lies under the boundaries for testing set. Another thing to consider is after it creates the boundaries of the data remaining training data becomes obsolete because we only need the core set of points which supports the boundaries to classify the new data set. This core data points called "support vectors". It is called vector because of each data point contains a row of observed data values for attributes [6].

[Figure 7 about here.]

Traditionally boundaries are called "hyperplanes" and it is used to describe boundaries in more than three dimensions because they are hard or sometimes impossible to visualize.[7]. Figure 7. Optimality of hyperplane expressed as a linear function which requires maximum distance between the identified classes. It only considers a small number of training example to build this hyperplane. SVM hyperplanes based on " separation of positive (+1) and negative (-1) with the largest margin [39]".

One of the main characteristic of the machine learning is to generalization. In other words, we want to give a general idea that tends to fit any of our testing datasets optimally. Support vector machines are a perfect regarding generalizations because once the training data fitted by the support vector machines other than support vector data inside the training data becomes redundant which means that even with the small changes inside the data will not have a significant effect on general boundaries [6].

Advantages of Support Vector Machines:

- Generalizes the data well with the help of boundaries. Which reduces the overfitting [6].

- Classification accuracy in basic support vector machine will yield a 95 percent accuracy with a default settings [6].
- SVM can deliver a unique solution, because of optimality solution is convex. This will give an advantage over Neural Networks which has multiple solutions in local minima [1].

Disadvantages of Support Vector Machines:

- One common disadvantage of SVM, is the lack of transparency because of its non-parametric techniques [1].
- Another biggest disadvantage of SVM is it requires high algorithmic complexity and high level of memory for the large-scale implementations [39].
- According to Burgees, biggest limitation of the SVM is in the choice of kernel [4].

5.8 Ensemble Methods

Ensemble methods goes into classification algorithm category, they are learning algorithms which uses weighted vote for it is prediction methods, in other words, it is learning rules over a small subset of data then we combine these rules which we learn from the small subset of data to make predictions and/or classification on the testing data [8]. The originality of the Ensemble method comes from Bayesian averaging, but with the recent algorithms include "Bagging, error-correcting, and boosting [8]".

Bagging refers to simply the looking at data-sets and dividing the data-set to it is small subsets then learning the rules of that particular small subset. Next step is combining each learned rule from subsets to apply to more significant data set. Combining method mostly done with averaging the learned rules. Bagging also does better on testing set than standard Linear Regression analysis and linear regression does better on training set especially in third order polynomial [8].

Stacking

Boosting is another method used in Ensemble Methods. The difference from bagging is in boosting we need to pick subsets or examples that we are not good at in other words hardest examples. Then we combine these learned rules with the weighted mean instead mean used in bagging method.

Boosting is little different then bagging.

Advantages of Ensemble Methods:

- Prediction of the ensemble methods is better than most of the algorithms because of the combining methods intuition makes the model less noisy [36].
- They are more stable than other algorithms. [36]

Disadvantages of Ensemble Methods:

- Over-fitting may cause some disadvantages for ensemble learning but bagging operation will reduce this overfitting [36].

6 FITTING DATA INTO MACHINE LEARNING ALGORITHMS

In this section, we will show the techniques we used on the execution of the prepared data into machine learning algorithms. Before fitting the data into the machine learning algorithms, we split the data into two sets. These sets are the training set and the testing set. We do splitting because of gaining an access of the future data will most likely be hard before future occurs, and because of this fact, it is a good idea to test our model with a dataset which our model has not seen it [40].

We used scikit-learn for splitting data into train and test we saved 20% of data for testing purposes as shown in Table 7 .

[Table 7 about here.]

Furthermore, after splitting the data we put all of our training data into to each of the machine learning algorithm to get their prediction results. We also provided code at the beginning and the end of each algorithm to calculate their running time.

Before we move further we need to discuss critical characteristics of a machine learning algorithm. These are;

- Confusion Matrix
- Accuracy
- Recall
- F-1 Score
- Precision

6.0.1 ***Confusion Matrix:*** Confusion matrix develops from 4 key elements. These elements are true positive, true negative, false negative, and false positive. As shown in Figure 8 about the constructing a confusion matrix. If we want to build a confusion matrix by targeting individuals who are making more than $50K our true positive, true negative, false positive, and false negative explained below.

[Figure 8 about here.]

True Positive (TP): We can explain true positive as if the individuals make more than $50K and our model correctly classifies them as individuals who makes more than $50K, then this individual is in higher income range, in this case, we call it a true positive [20].

True Negative (TN): Intuition of true negative is if an individual makes less than $50K and our model correctly classifies them as individuals who makes less then $50K, then this individual is in lower income range. We call this true negative [20].

False Negative (FN): When an individual makes less than $50K and our model incorrectly classifies them in higher income range by making a mistake causes a false negative to happen [20].

False Positive (FP): When an individual is making more than $50K and our model classifies them in lower income range by mistake. This is called false positive [20].

6.0.2 ***Accuracy:*** Accuracy answers the question of how good is the model is. In our case this question will be out of all the individuals, how many did the models classify the individuals correctly. The mathematical expression of the accuracy is the ratio between the number of correctly classified points and the number of total points. We can think that if we have high accuracy, our model is excellent, but this is only where we have identical false positive and false negative values in our dataset [20].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

6.0.3 ***Precision***. Precision answers the questions of out of all the points predicted to be positive how many of them were actually positive? If we translate this question into our case, we will have

out of all the individuals that we are classified as lower income how many were actually have lower income. Higher precision indicates that we have low false positive rate [20]. Mathematical expression of precision is;

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

6.0.4 ***Recall (Sensitivity)***. Recall answers the question of "out of the points that are labeled positive how many of them were correctly predicted is positive ? ". If we translate this to our case, we will have "out of the points that are labeled higher income how many of them correctly predicted is in higher income range ? ". Mathematical expression of the recall is;

$$Precision = \frac{TP}{TP + FN} \tag{5}$$

6.0.5 ***F-1 Score***. The F-1 score is the idea of giving a decision by looking at only one score which will include precision, and recall scores. We cannot just take the average of precision and recall because if either of them is very low. We need a number to be low, even id the other one is not. This will leads us to look at the harmonic mean, and it works as follow. Let's say we have two numbers X and Y. X is smaller than Y, and we have the arithmetic mean, and it always lies between X and Y. It is a mathematical fact that the harmonic mean is always less than the arithmetic mean which is closer to the smaller number than to the higher number. Mathematical expression of F-1 score is;

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

6.1 Results

Now we can look at the results from each of the machine learning algorithm. Results also showed in Table 8 with the visualization of Figure 10. We can also see the running time of the each of the algorithm in Figure 9. Support Vector Machines is the winner for the highest running time for training the algorithm.

[Figure 9 about here.]

[Table 8 about here.]

6.1.1 ***Naive Bayes***. As shown in the Figure 10 we have a comparison of several supervised machine learning algorithms on our dataset. We can see that from the accuracy standpoint Naive Bayes algorithms have the lowest score which means that it did not do a good job for labeling true positives regards to all data but it did a good job in precision standpoint while doing a bad classification from recall standpoint. Two key element for us in this situation is accuracy and f1 score(which consist of precision and recall).

6.1.2 ***Support Vector Machine***. Support Vector Machine is the second best algorithm in our case. This algorithm did very well job on classification it has the second highest accuracy and f1 score.

6.1.3 ***AdaBoost***. As we stated before ensemble algorithms learn from the small portion of the data and combine these learning to do the predictive task. As shown in Figure 10 adaboosting has the highest accuracy score among all the other algorithms. This algorithm should be our first choice to do predictive modeling. We believe that there is still an improvements on accuracy

6.1.4 ***K-Nearest Neighboors***. K-Nearest Neighbor algorithm in our project we set the k value to 5. K Nearest Neighbor algorithm also did a good job by placing itself third in accuracy score.

6.1.5 ***Decision Tree***. Decision Tree is gave a good accuracy but fall behind on f1 score as shown in Figure 10.

[Figure 10 about here.]

7 CONCLUSION

We presented the importance of analytical approach with machine learning algorithms and how they can be used to predict or classify the individuals with many different attributes like age, education, income, etc. We also presented weaknesses and strengths of these algorithms along with their precision, accuracy, recall, and F-1 scores by presenting with the visualizations. We also demonstrated the running time for each algorithm while using big data sets. The source code of this project can found Github website which presented in reference section [44].

ACKNOWLEDGMENTS

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

REFERENCES

[1] L. Auria and A. R. Moro. 2008. Support Vector Machines (SVM) as a Technique for Solvency Analysis. Online. http://www.diw.de/english/products/publications/discussion_papers/27539.html

[2] L. Ben. 2015. Six Reasons why I recommend scikit-learn. Online. (Oct. 2015). https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn

[3] L. Bottou. 2010. Stochastic Gradient Descent. Online. (2010). http://leon.bottou.org/projects/sgd

[4] C. J. C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 2 (01 Jun 1998), 121–167. https://doi.org/10.1023/A:1009715923555

[5] N. Deanna, S. Nathan, and W. Rachel. 2016. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming* 155, 1 (01 Jan 2016), 549–573. https://doi.org/10.1007/s10107-015-0864-7

[6] B. Deshpande. 2013. When do support vector machines trump other classification methods. Online. (Jan. 2013). http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-other-classification-methods

[7] B. Desphpande. 2011. 4 key advanatages of using decision trees for predictive analytics. Online. (July 2011). http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics

[8] G. T. Dietterich. n.d.. Ensemble Methods in Machine Learning. (n.d.). http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

[9] EliteDataScience. 2016. Modern Machine Learning Algorithms: Strengths and Weaknesses. Online. (May 2016). https://elitedatascience.com/machine-learning-algorithms

[10] O. F. Ertugrul and M. E. Tagluk. 2017. A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing* 55, Supplement C (2017), 480 – 490. https://doi.org/10.1016/j.asoc.2017.02.020

[11] M. Fan. n.d.. How and Why to Use Stochastic Gradient Descent? (n.d.). http://anson.ucdavis.edu/~minjay/SGD.pdf

[12] J. Fang. 2013. Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses. *Journal of Business and Economics* 4, 7 (July 2013), 620–633. http://www.academicstar.us/UploadFile/Picture/2014-6/201461494819669.pdf

[13] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem. 2017. Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renewable Energy* 111, Supplement C (2017), 52 – 62. https://doi.org/10.1016/j.renene.2017.03.083

[14] S. T. Indra, L. Wikarsa, and R. Turang. 2016. Using logistic regression method to classify tweets into the selected topics. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on* 1, 385-389 (2016), 385. http://proxyiub.uits.iu.edu/login?url=https:

//search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edseee&AN=edseee.7872727&site=eds-live&scope=site
[15] Investopedia. n.d.. Correlation Coefficient. Online. (n.d.). https://www.investopedia.com/terms/c/correlationcoefficient.asp
[16] D. S. Jadhav and H. P. Channe. 2014. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)* 5, 1 (Jan. 2014), 1842–1845. https://www.ijsr.net/archive/v5i1/NOV153131.pdf
[17] B. Jason. 2014. A gentle introduction to Scikit-Learn: Python Machine Learning Library. Online. (April 2014). https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/
[18] H. Jeff. 2012. Introduction to Machine Learning. Online. (Jan. 2012). http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf
[19] J. Jiaqi and Y. Chung. 2017. Research on K nearest neighbor join for big data. In *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE, Department of Computer Engineering Wonkwang University Iksan 54538, Korean, 1077–1081. https://doi.org/10.1109/ICInfA.2017.8079062
[20] R. Joshi. 2016. Accuracy, Precision, Recall, and F1 Score: Interpretation of Performance Measures. Online. (Sept. 2016). http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures
[21] R. Kohavi. 1996. Improving the Accuracy of Naive-Bayes Classifiers: A Decision-tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Silicon Graphics, Inc, 202–207. http://dl.acm.org/citation.cfm?id=3001460.3001502
[22] R. Kohavi and B. Becker. n.d.. Predicting whether income exceeds $50K/yr based on census data. Online. (n.d.). https://archive.ics.uci.edu/ml/datasets/Census+Income
[23] J. Kunal. 2015. Scikit-Learn in python - The most important Machine Learnig Tool I learnt last year. Online. (Jan. 2015). https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/
[24] M. D. Lane. n.d.. Log Transformations. Online. (n.d.). http://onlinestatbook.com/2/transformations/log.html
[25] V. Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng. 2011. On optimization methods for deep learning. In *International Conference of Machine Learning*. Stanford University, International Conferenfe of Machine Learning, Stanford University, NA. https://cs.stanford.edu/~acoates/papers/LeNgiCoaLahProNg11.pdf
[26] Pandas Library. n.d.. Dataframe replace. Online. (n.d.). https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html
[27] Pandas Library. n.d.. Pandas Dateframe describe. Online. (n.d.). https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html
[28] Pandas Py Data Library. n.d.. Pandas for Python. Online. (n.d.). https://pandas.pydata.org/
[29] L. J. Moon. 2017. Fast k-Nearest Neighbor Searching in Static Objects. *Wireless Personal Communications* 93, 1 (01 Mar 2017), 147–160. https://doi.org/10.1007/s11277-016-3524-1
[30] G. Nick. 2014. KNN. Online. (April 2014). http://www.nickgillian.com/wiki/pmwiki.php/GRT/KNN
[31] R. Nick. NA. The Disadvantages of Logistic Regression. Online. (NA). http://classroom.synonym.com/disadvantages-logistic-regression-8574447.html
[32] C. Petri. 2010. Decison Trees. Online. (2010). http://www.cs.ubbcluj.ro/~gabis/DocDiplome/DT/DecisionTrees.pdf
[33] U. Princeton. NA. Decision Tree Learning. Online. (NA). http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf
[34] S. A. Raj, L. J. Fernando, and S. Raj. 2017. Predictive Analytics On Political Data. Congress. *World Congress on Computing and Communication Technologies* 10, 1109 (2017), 93–96.
[35] M. Ray. 2012. Nearest Neighbours: Pros and Cons. Online. (April 2012). http://www2.cs.man.ac.uk/~raym8/comp37212/main/node264.html
[36] S. Ray. 2015. 5 Easy Questions on Ensemble Modeling Everyone Should Know. Online. (Jan. 2015). https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/
[37] J. Rie and Z. Tong. 2013. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Rutgers University, New Jersey, USA, 315–323. http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf
[38] Scikitlearn. n.d.. Stochastic Gradient Descent. Online. (n.d.).
[39] K. N. Shrivastava, P. Saurabh, and B. Verma. 2011. An Efficient Approach Parallel Support Vector Machine for Classification of Diabetes Dataset. *International Journal of Computer Applications in Technology* 36, 6 (Dec. 2011), 19–24. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.3757&rep=rep1&type=pdf
[40] D. Steinberg. 2014. Why Data Scientist Split Data into Train and Test. Online. (March 2014). https://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test
[41] K. B. Tapan. 2015. Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *Inteligencia Artificial, Vol 18, Iss 56, Pp 14-30 (2015)* 1, 56 (2015), 14. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edsdoj&AN=edsdoj.0e372b34c5d48bcb72cd437eede1fd1&site=eds-live&scope=site
[42] A. F. Tehrani, W. Cheng, and E. Hullermeier. 2011. Choquistic Regression: Generalizing Logistic Regression Using the Choquet Integral. Online. (July 2011). https://www-old.cs.uni-paderborn.de/fileadmin/Informatik/eim-i-is/PDFs/Talk.EUSFLAT.11.pdf
[43] K. Teknomo. 2017. K-Nearest Neighbor Tutorial. Online. (2017). http://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm
[44] E. B. Usifo. 2017. Income Prediction. Github. (Dec. 2017). https://github.com/bigdata-i523/hid343/tree/master/project
[45] R. Vasudev. n.d.. What is One Hot Encoding? do you have to use it ? Online. (Aug. n.d.). https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f
[46] Wikipedia. 2017. Naive Bayes. Online. (Nov. 2017). https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[47] Wikipedia. NA. Feauture Scaling. Online. (NA). https://en.wikipedia.org/wiki/Feature_scaling
[48] Wikipedia. n.d.. Comma Seperated Values. Online. (n.d.). https://en.wikipedia.org/wiki/Comma-separated_values
[49] Wikipedia. n.d.. Decision Trees. Online. (n.d.). https://en.wikipedia.org/wiki/Decision_tree
[50] H. Zhang. 2004. *The Optimality of Naive Bayes*. resreport. University of New Brunswick. http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

List of Figures

1 Scatter Matrix Plot [44]. 9
2 Histogram of Capital Gain [44]. 10
3 Histogram of Capital Loss [44]. 10
4 After Logarithmic Function Applied Histogram of Capital Gain [44]. 11
5 Example of Naive Bayes [50]. 12
6 Example of Decision Tree Construction[33]. 12
7 Example of Shows the Hyperplanes [6]. 13
8 Example of Confusion Matrix Construction [20]. 13
9 Supervised Learning Algorithm Running Time Results [44]. 14
10 Supervised Learning Algorithm Results [44]. 15

Figure 1: Scatter Matrix Plot [44].

Figure 2: Histogram of Capital Gain [44].

Figure 3: Histogram of Capital Loss [44].

Figure 4: After Logarithmic Function Applied Histogram of Capital Gain [44].

Figure 5: Example of Naive Bayes [50].

Figure 6: Example of Decision Tree Construction[33].

Figure 7: Example of Shows the Hyperplanes [6].

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 8: Example of Confusion Matrix Construction [20].

Figure 9: Supervised Learning Algorithm Running Time Results [44].

Figure 10: Supervised Learning Algorithm Results [44].

List of Tables

1 Statistical Summary of The Continuous Variables 17
2 Description of the Binary Values 17
3 Example of One Hot Encoding [45]. 17
4 Example of One Hot Encoding After [45]. 17
5 Statistical Summary of Continuous Variables [44]. 17
6 Correlation Matrix [44]. 18
7 Train-Test-Split [44]. 18
8 Results of the Algorithms [44]. 18

	age	education	cap gain	cap loss	hours
count	32561	32561	32561	32561	32561
mean	38.581	10.08	1077.64	87.303	40.437
std.	13.640	2.572	7385.292	402.960	12.347
min.	17.0	1.0	0	0	1.0
25%	28.0	9.0	0	0	40.0
50%	37.0	10.0	0	0	40.0
75%	48.0	12.0	0	0	45.0
max	90.0	16.0	0	4356.0	99.0

Table 1: Statistical Summary of The Continuous Variables

Description	Assigned Value
Individuals who makes more than $50K	1
Individuals who makes at or less than $50K	0

Table 2: Description of the Binary Values

Description	Assigned Value
Individuals who makes more than $50K	1
Individuals who makes at or less than $50K	0

Table 3: Example of One Hot Encoding [45].

VW	Acura	Honda	Price
1	0	0	20,000
0	1	0	10,011
0	0	1	50,000
0	0	1	10,000

Table 4: Example of One Hot Encoding After [45].

	Age	Gain	Loss	Hours
Number of Instances	32,561	32,561	32,561	32,561
Mean	38.58	1077.64	87.303	40.437
Standard Deviation	13.640	7385.292	402.960	12.347
Minimum Value	17	0	0	1
25th percentile	28	0	0	40
50th percentile	37	0	0	40
75th percentile	48	0	0	45
Maximum Values	90	99999	4356	99

Table 5: Statistical Summary of Continuous Variables [44].

	Age	**Education**	**Capital Gain**	**Capital Loss**	**Hours Per Week**
Age	1.0	0.043	0.080	0.060	0.101
Education	0.043	1.0	0.124	0.079	0.152
Capital Gain	0.080	0.124	1.0	-0.032	0.080
Capital Loss	0.060	0.796	-0.032	1.0	0.052
Hours Per Week	0.101	0.152	0.080	0.052	1.0

Table 6: Correlation Matrix [44].

Splitting the Data	**Sample Size**
Training	24129
Testing	6033

Table 7: Train-Test-Split [44].

Name	Accuracy	Recall	Precision	F1 Score
Naive Bayes	0.4442	0.4642	0.9680	0.3053
SVC	0.8301	0.5969	0.5056	0.7284
AdaBoost	0.8499	0.6724	0.6189	0.7361
KNN	0.8184	0.6090	0.5682	0.6561
Decision Tree	0.8161	0.6231	0.6109	0.6459

Table 8: Results of the Algorithms [44].

```
bibtext report
===========================================

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

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

report.bib:518:@incollection{NIPS2013_4937,

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

report.bib:172:@INPROCEEDINGS{knn-chung,

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


latex report
===========================================

        [2017-12-04 22.02.49] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 2.0s.
../README.yml
  8:81      error    line too long (188 > 80 characters)  (line-length)
  8:188     error    trailing spaces  (trailing-spaces)
  21:81     error    line too long (534 > 80 characters)  (line-length)
  34:81     error    line too long (666 > 80 characters)  (line-length)
  34:666    error    trailing spaces  (trailing-spaces)
  35:12     error    trailing spaces  (trailing-spaces)
  37:30     error    trailing spaces  (trailing-spaces)
  42:5      error    duplication of key "type" in mapping  (key-duplicates)


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

```
name: Usifo, Borga
hid:  343
paper1: 100 %
paper2: 100 %
project: 100 %

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

18
wc 343 project 18 5776 report.tex
wc 343 project 18 6315 report.pdf
wc 343 project 18 3252 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
--------------------------------------------------------------------------------

47: We first used the pandas \cite{www-pandas} to help to load the
```

data in data frame format. This gave us a unique advantage, and faster processing of comma separated values for putting into data frame \cite{www-commasep}. Our data consist of 15 variables. Some of these variables are continuous, and some of them are categorical variables, and our target variable was ``income`` attribute. After putting the data into data frames, we first got a statistical snapshot of continuous variables (age, education, capital gain, capital loss, hours worked) by using the pandas \cite{www-pandas.describe} functions as shown in Table \ref{stats-table}.

49: \begin{table}[!ht]

64: \label{stats-table}

76: \par Our last process is changing the string values to numerical values on our target data which consist of string values (``\$ 50K``) for machine learning algorithms to understand this target data we need to transfer it to numerical values. Since we have only two categories, we will assign 1 and 0 as numerical values as shown in Table \ref{assign-values}.

78: \begin{table}[!ht]

87: \label{assign-values}

90: \par Our shape of the data will also receive impact from changing to numerical. Our number of futures will go from 14 to 103. This is because we implemented one-hot-encode to our dataset. It is called one hot encoded because we transform the categorical variables into a more acceptable shape for the machine learning algorithms to perform well \cite{www-hackernoon}. In other words ``we implement binarization of the category to include as a future to train model \cite{www-hackernoon}``. As we can see in Table \ref{one-hot-before} and Table \ref{one-hot-after}.

92: \begin{table}[!ht]

101: \label{one-hot-before}

104: \begin{table}[!ht]

115: \label{one-hot-after}

121: After cleaning the data, we started our data exploration to learn little bit more from our data and make necessary changes if needed before putting into our machine learning algorithms. The first step in this process is getting the total count of the individuals as well as the count of the individuals who are making more than \$50K and less than \$50K which can be seen in below Table \ref{my-label-2}.

132: \label{my-label-2}

135: \par Moreover, we also look at the statistical values of each of the continuous variable we have. Those values given in Table \ref{my-label}. As we can see we have individuals who're age ranging from 17 to 90 years old with a mean of 38.58. If we look at the capital gains and capital losses, we have a standard

```
deviation of 7385 and 402 respectively this is also another
indication of skew in these variables.
137: \begin{table}[!ht]
152: \label{my-label}
155: \par We used scatter matrix plot and applied the correlation
function to see if we have any reliable correlation between any
of the variables. As we can see from the correlation matrix Table
\ref{scatter-matrix} and correlation numbers Figure
\ref{fig:scatter}we do not have the high correlation between any
variables. Correlation values range between -1 to 1. The
correlation value of 1 is an indication of perfect positive
correlation and correlation number -1 indicates a negative
correlation between variables \cite{www-investopedia}. Because of
lower correlation values, it will be tough to determine the
classification by just looking at the correlations; this
indicates we have sophisticated algorithms to determine the
relationship between variables to classify individuals incomes.
157: \begin{table}[!ht]
169: \label{scatter-matrix}
173: \begin{figure}[!ht]
175: \includegraphics[width=\columnwidth]{images/scatter-matrix.png}
176: \caption{Scatter Matrix Plot
\cite{Borga2017}.}\label{fig:scatter}
179: \par Furthermore, we also explore the capital gains, capital
losses, and hours per week variables which we used a histogram to
plot the data into distribution form so we can see how all these
attributes distributed. The reason we do the histogram is we want
to see any skewness in our data. As shown in the histogram graphs
in Figure \ref{fig:Hist-capital} and  Figure \ref{fig:loss-
capital} in capital gains and capital loss we have highly skewed
data which can cause issues later on in our algorithms. We apply
a logarithmic function to do highly skewed data to less skewed
\cite{www-onlinestat}. Using logarithmic functions adds more
value to data from the interpretable standpoint and ``it helps to
meet the assumptions of inferential statistics \cite{www-
onlinestat}``.
181: \begin{figure}[!ht]
183: \includegraphics[width=\columnwidth]{images/capital-gain.png}
184: \caption{Histogram of Capital Gain
\cite{Borga2017}.}\label{fig:Hist-capital}
187: \begin{figure}[!ht]
189: \includegraphics[width=\columnwidth]{images/capital-loss.png}
190: \caption{Histogram of Capital Loss
\cite{Borga2017}.}\label{fig:loss-capital}
193: \par Moreover, applying logarithmic function had an impact on
distribution. We can see the changes on skew data in Figure
```

```
    \ref{fig:Hist-capital-log} after applying logarithmic function.
195: \begin{figure}[!ht]
197: \includegraphics[width=\columnwidth]{images/logarithmic-
    applied.png}
198: \caption{After Logarithmic Function Applied Histogram of Capital
    Gain \cite{Borga2017}.}\label{fig:Hist-capital-log}
235: \par Term naive comes from it is the method of use probability
    among categories which assumes of independence among given class
    of attributes as shown in Figure \ref{fig:Naive Bayes}. In other
    words, if we try to classify individuals from their email
    communications it will not take the order of words into account.
    Whereas in the English language we can tell the difference
    between sentence makes sense or not if we randomly re-order our
    words in the sentences. So it does not understand the text, it
    only looks at word frequencies as a way to do the classification.
    This is why it is called ``Naive``.
237: \begin{figure}[!ht]
240: \includegraphics[width=\columnwidth]{Naive-bayes}
241: \caption{Example of Naive Bayes \cite{Zhang}.}\label{fig:Naive
    Bayes}
326: \par Representation of decision trees is as follows; we sort the
    instances from root to leaf nodes, this sorting gives insights
    about the classification of the instance, every outcome
    descending from the root node corresponds to possible values for
    that variable \cite{www-cs.princeton}. We can classify an
    instance by starting from the root node and checking the
    attributes labeled on that node and moving down from that node
    based on attribute given attribute values \cite{www-cs.princeton}
    as shown in Figure \ref{fig:Decision Tree}.
328: \begin{figure}[!ht]
330: \includegraphics[width=\columnwidth]{images/decison_tree.png}
331: \caption{Example of Decision Tree Construction\cite{www-
    cs.princeton}.}\label{fig:Decision Tree}
379: \begin{figure}[!ht]
381: \includegraphics[width=\columnwidth]{images/hyperplane-
    boundary.png}
382: \caption{Example of Shows the Hyperplanes \cite{www-simafore-
    svm}.}\label{fig:Hyperplane}
385: \par Traditionally boundaries are called ``hyperplanes`` and it
    is used to describe boundaries in more than three dimensions
    because they are hard or sometimes impossible to
    visualize.\cite{www-simafore}. Figure \ref{fig:Hyperplane}.
    Optimality of hyperplane expressed as a linear function which
    requires maximum distance between the identified classes. It only
    considers a small number of training example to build this
    hyperplane. SVM hyperplanes based on `` separation of positive
```

```
(+1) and negative (-1) with the largest margin \cite{verma-
ssv}``.
430: \par We used scikit-learn for splitting data into train and test
we saved 20\% of data for testing purposes as shown in Table
\ref{split} .
432: \begin{table}[!ht]
441: \label{split}
457: Confusion matrix develops from 4 key elements. These elements are
true positive, true negative, false negative, and false positive.
As shown in Figure \ref{fig:confusion-matrix} about the
constructing a confusion matrix. If we want to build a confusion
matrix by targeting individuals who are making more than \$50K
our true positive, true negative, false positive, and false
negative explained below.
459: \begin{figure}[!ht]
461: \includegraphics[width=\columnwidth]{images/confusion-matrix.png}
462: \caption{Example of Confusion Matrix Construction \cite{www-
exsilio}.}\label{fig:confusion-matrix}
504: Now we can look at the results from each of the machine learning
algorithm. Results also showed in Table \ref{result-table} with
the visualization of Figure \ref{fig:result-algo}. We can also
see the running time of the each of the algorithm in Figure
\ref{fig:result-time}. Support Vector Machines is the winner for
the highest running time for training the algorithm.
506: \begin{figure}[!ht]
508: \includegraphics[width=\columnwidth]{images/running-time.png}
509: \caption{Supervised Learning Algorithm Running Time Results
\cite{Borga2017}.}\label{fig:result-time}
512: \begin{table}[!ht]
524: \label{result-table}
528: As shown in the Figure \ref{fig:result-algo} we have a comparison
of several supervised machine learning algorithms on our dataset.
We can see that from the accuracy standpoint Naive Bayes
algorithms have the lowest score which means that it did not do a
good job for labeling true positives regards to all data but it
did a good job in precision standpoint while doing a bad
classification from recall standpoint. Two key element for us in
this situation is accuracy and f1 score(which consist of
precision and recall).
532: As we stated before ensemble algorithms learn from the small
portion of the data and combine these learning to do the
predictive task. As shown in Figure \ref{fig:result-algo}
adaboosting has the highest accuracy score among all the other
algorithms. This algorithm should be our first choice to do
predictive modeling. We believe that there is still an
improvements on accuracy
```

```
537: Decision Tree is gave a good accuracy but fall behind on f1 score
     as shown in Figure \ref{fig:result-algo}.
544: \begin{figure}[!ht]
546: \includegraphics[width=\columnwidth]{images/result-score.png}
547: \caption{Supervised Learning Algorithm Results
     \cite{Borga2017}.}\label{fig:result-algo}

figures 10
tables 8
includegraphics 10
labels 19
refs 17
floats 18

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

Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------


WARNING: algorithm and below may be used improperly

121: After cleaning the data, we started our data exploration to learn
     little bit more from our data and make necessary changes if
     needed before putting into our machine learning algorithms. The
     first step in this process is getting the total count of the
     individuals as well as the count of the individuals who are
     making more than \$50K and less than \$50K which can be seen in
     below Table \ref{my-label-2}.
```

```
WARNING: code and below may be used improperly

211: Scikit-learn developed by David Cournapeau in 2007. The
     development came from while he was working on summer code project
     for Google. After recognized and published by INRIA in 2010
     project start the get more attention among worldwide. There are
     more than 30 active contributors and has secured several
     sponsorships from big technology companies\cite{www-
     machinelearningmystery}. ``It also has a goal of providing common
     algorithms to Python users through consistent interface\cite{www-
     oreily}``. Scikit-Learn consists of several elements to make
     analytical predictions. These elements are shown below\cite{www-
     analyticvidhya}:

WARNING: algorithm and below may be used improperly

211: Scikit-learn developed by David Cournapeau in 2007. The
     development came from while he was working on summer code project
     for Google. After recognized and published by INRIA in 2010
     project start the get more attention among worldwide. There are
     more than 30 active contributors and has secured several
     sponsorships from big technology companies\cite{www-
     machinelearningmystery}. ``It also has a goal of providing common
     algorithms to Python users through consistent interface\cite{www-
     oreily}``. Scikit-Learn consists of several elements to make
     analytical predictions. These elements are shown below\cite{www-
     analyticvidhya}:

WARNING: algorithm and below may be used improperly

221: \par Algorithms that we will investigate shown below; we will go
     more deep analysis on each of these algorithms.

bibtex
--------------------------------------------------------------------------------


label errors

518: NIPS2013_4937: do not use underscore in labels:

bibtex errors

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

```
Database file #1: report.bib

bibtex_empty_fields
--------------------------------------------------------------------------------


entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------


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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

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 INTRODUCTION

Part of the kaggle competition, two sample data sets are given with 80 attributes (variables) describing various aspects of the residential homes in Ames and Iowa cities. Training dataset 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 kaggle competition and achieved better kaggle scores. Kaggle score is a measure to indicate accuracy and the quality of the algorithm. We have applied various exploratory analysis techniques and engineer the features before applying a few advanced supervised learning algorithms.

2 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 attribute details are given in the appendix section as a reference.

2.1 Handling Missing Values

First part of the analysis, we have checked for the missing values in the dataset. we have also identified that there are 6 variables having most of the missing data. All the missing values for the numeric variables are analyzed further to decide whether to delete the instances of the data with missing values or fill them using meaningful data such as median of the corresponding variable.

```
# 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 about here.]

2.2 Analyze Numerical Variables

There are 37 numeric variables after excluding the *Id* variable. We have analyzed the numeric variables for data patterns such as skewed data and range of the possible values. *over all quality*, *ground living area*, *garage cars* and *garage area* graphs are included as samples.

```
# 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 2 about here.]

[Figure 3 about here.]

2.3 Analyze Categorical Variables

There are 43 categorical variables in the dataset. We have analyzed all categorical variables and found the ways to fill the missing values and also the way we need to convert them into numerical variables. Later in the feature engineering section, we will go through the details. *neighborhood* and *sale type* graphs are included as samples.

```
# 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 4 about here.]

2.4 Analyze Correlations

Numpy package offers correlations functionality to analyze the variables that highly positively or negatively correlated with *sale price* and with the other variables. We can visualize a few pair-wise correlation graphs with sale price for detailed analysis. From the correlation-plot we can list the top 10 features those are strongly correlated with the target variable - *sale price*.

```
# 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 5 about here.]

[Figure 6 about here.]

[Figure 7 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

2.5 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.

2.6 Outlier Analysis

Continuing with exploratory analysis, We have analyzed the outliers using *Cook's distance* and then further analyzed two key variables - *ground live area* and *garage area* that are in high correlation with *sale price*. We have removed the outlier rows related to these two variables as they are only 8 rows impacting the training dataset.

```
# 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 8 about here.]

```
# 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 9 about here.]

[Figure 10 about here.]

2.7 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 featues 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. 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 data set. 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 category 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 the dimensions in the model. After extensive feature engineering applied on the housing dataset, we have added *228* new features (variables). Following are the python methods to factorize categorical variables and one-hot encoding techniques.

```
# 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
```

3 ALGORITHMS AND METHODOLOGY

Linear regression predicts the target variable using best possible straight line fit to the set predictor variables. The possible 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, XGBoost and Neural Network, to predict the test data housing prices.

3.1 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. Following python code snippet is the implementation for *sale price* predictions of the housing test dataset.

```
# 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)
```

3.2 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. Following are the random forest algorithm details for the *sale price* predictions implemented in python using *sklearn* package.

```
# 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 11 about here.]

3.3 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. Following is the lasso algorithm implementation for *sale price* predictions in python.

```
# 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)
```

3.4 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, where as L2 adds the penalty equivalent to *square of the magnitude* of the variable coefficients. Following is the python implementation of the ridge algorithm for the *sale price* predictions.

```
# 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 12 about here.]

[Figure 13 about here.]

3.5 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 *Fscore* to measure the importance of variables. Following table explains the hyper-parameters of XGBoost algorithm and also given the python code implementing XGBoost algorithm for *sale price* predictions.

[Table 1 about here.]

```
# 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 14 about here.]

3.6 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.

3.7 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. Following are the kaggle public scores that we have achieved by submitting all the six algorithms along with the ensembled model. 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 15% of the competition. Following table displays each algorithm and the *root-mean-square error* (RMSE) along with the *kaggle score*.

[Table 2 about here.]

4 DEVELOPMENT ENVIRONMENT

- Operating Environment: Ubuntu 16.4 through Oracle Virtual Box 5.2
- Programming Language: Python 2.7
- Development Tools/Environment: Jupyter Notebook, Anaconda (data science platform)
- Python Libraries: numpy, pandas, sklearn, matplotlib, seaborn, xgboost and tensorflow
- Repository: git@github.com:bigdata-i523/hid306.git
- Project Folders:
  - Code: all Jupyter notebook files
  - Images: all output images as PDF files
  - Data: all the input and output datasets in CSV format
- Python Jupyter notebook files:
  - 1.1_exploratory_analysis_numerical.ipynb
  - 1.2_exploratory_analysis_categorical.ipynb
  - 1.3_outlier_and_skewed_data_analysis.ipynb
  - 1.4_feature_engineering.ipynb
  - 2.1_algorithm_svm.ipynb
  - 2.2_algorithm_random_forest.ipynb
  - 2.3_algorithm_ridge.ipynb
  - 2.4_algorithm_lasso.ipynb
  - 2.5_algorithm_neural_network_tf.ipynb
  - 2.6_algorithm_xgboost.ipynb
  - 3_ensemble_kaggle_submission.ipynb
- Input Datasets:
  - train.csv
  - test.csv
- Kaggle Submissions:
  - kaggle_python_svm.csv

- kaggle_python_random_forest.csv
- kaggle_python_ridge.csv
- kaggle_python_xgboost.csv
- kaggle_python_lasso.csv
- kaggle_python_neural_network.csv
- kaggle_python_ensemble.csv

5 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 KAGGLE HOUSING PRICE DATASET VARIABLES

- 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

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/notaapple/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 XGBoost. (March 2016). https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[5] Selva Prabhakaran. [n. d.]. Outlier Treatment. ([n. d.]). 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 Variables with Missing Values 9
2 Sample Numerical Variables 10
3 Sample Category Variables 10
4 Correlations 11
5 Pair-wise Correlations with Sale Price 12
6 Outliers - Cook's distance 12
7 Outliers - Garage Live Area 13
8 Outliers - Garage Area 13
9 Random Forest Feature Ranking 14
10 Ridge Algorithm - Top 10 Positive Features 15
11 Ridge Algorithm - Top 10 Negative Features 16
12 XGBoost Feature Importance 17

Figure 1: Variables with Missing Values

9

Figure 2: Sample Numerical Variables

Figure 3: Sample Category Variables

Figure 4: Correlations

11

Figure 5: Pair-wise Correlations with Sale Price

Figure 6: Outliers - Cook's distance

Figure 7: Outliers - Garage Live Area

Figure 8: Outliers - Garage Area

Figure 9: Random Forest Feature Ranking

Figure 10: Ridge Algorithm - Top 10 Positive Features

Figure 11: Ridge Algorithm - Top 10 Negative Features

Figure 12: XGBoost Feature Importance

List of Tables

1 XGBoost Hyper-parameters 19
2 Kaggle Submissions 19

Table 1: 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 iteratio.n

Table 2: 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
Warning--empty year in 1
(There was 1 warning)

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

report.bib:59:@Misc{Aarshay_Jain2016,

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


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


latex report
===========================================

        [2017-12-04 22.01.32] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
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 1.0s.
../README.yml
  41:1      error    trailing spaces  (trailing-spaces)
  45:62     error    trailing spaces  (trailing-spaces)
  46:67     error    trailing spaces  (trailing-spaces)
  47:75     error    trailing spaces  (trailing-spaces)
  48:69     error    trailing spaces  (trailing-spaces)
  49:64     error    trailing spaces  (trailing-spaces)
  53:26     error    trailing spaces  (trailing-spaces)
  58:28     error    trailing spaces  (trailing-spaces)


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

```
================================================================================


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

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

19
wc 306 project 19 3869 report.tex
wc 306 project 19 4279 report.pdf
wc 306 project 19 197 report.bib

find "
--------------------------------------------------------------------------------

49: train = pd.read_csv("../data/train.csv")

50: test = pd.read_csv("../data/test.csv")

130: sns_plot = sns.heatmap(corr, cmap="YlGnBu",

177: model = ols(formula = "SalePrice ~

178: GrLivArea + GarageArea", data=train)

181: criterion="cooks")

224: dummies = pd.get_dummies(df[col_name], prefix="_" + col_name)

258: train = pd.read_csv("train.csv")

259: test = pd.read_csv("test.csv")

260: target_vector = train["SalePrice"]

288: train = pd.read_csv("train.csv")
```

```
289: test = pd.read_csv("test.csv")

290: target_vector = train["SalePrice"]

326: train = pd.read_csv("train.csv")

327: test = pd.read_csv("test.csv")

328: target_vector = train["SalePrice"]

358: train = pd.read_csv("train.csv")

359: test = pd.read_csv("test.csv")

360: target_vector = train["SalePrice"]

393: df_predict = pd.DataFrame({'Id': test["Id"],

446: train = pd.read_csv("train.csv")

447: test = pd.read_csv("test.csv")

448: target_vector = train["SalePrice"]

passed: False

find footnote
--------------------------------------------------------------------------------

passed: True

find input{format/i523}
--------------------------------------------------------------------------------

4: \input{format/i523}

passed: True

find input{format/final}
--------------------------------------------------------------------------------

passed: False

floats
--------------------------------------------------------------------------------
```

```
59: \begin{figure}[H]
61: \includegraphics[width=0.9\columnwidth]{images/missing_values}
62: \caption{Variables with Missing Values} \label{fig:missing_values}
79: \begin{figure}[H]
81: \includegraphics[width=0.75\columnwidth]{images/num_features_1}
84: \begin{figure}[H]
86: \includegraphics[width=0.75\columnwidth]{images/num_features_2}
87: \caption{Sample Numerical Variables} \label{fig:num_features_2}
114: \begin{figure}[H]
116: \includegraphics[width=1.0\columnwidth]{images/cat_features_1}
117: \caption{Sample Category Variables} \label{fig:cat_features_1}
134: \begin{figure}[H]
136: \includegraphics[width=1.2\columnwidth]{images/correlations}
137: \caption{Correlations} \label{fig:correlations}
140: \begin{figure}[H]
142: \includegraphics[width=0.9\columnwidth]{images/pair_wise_correlat
     ions_1}
145: \begin{figure}[H]
147: \includegraphics[width=0.9\columnwidth]{images/pair_wise_correlat
     ions_2}
148: \caption{Pair-wise Correlations with Sale Price}
     \label{fig:pair_wise_correlations_2}
184: \begin{figure}[H]
186: \includegraphics[width=.95\columnwidth]{images/outliers}
187: \caption{Outliers - Cook's distance} \label{fig:outliers}
199: \begin{figure}[H]
201: \includegraphics[width=.95\columnwidth]{images/gr_liv_area_outlie
     r}
202: \caption{Outliers - Garage Live Area}
     \label{fig:gr_liv_area_outlier}
205: \begin{figure}[H]
207: \includegraphics[width=.95\columnwidth]{images/garage_area_outlie
     r}
208: \caption{Outliers - Garage Area} \label{fig:garage_area_outlier}
311: \begin{figure}[H]
313: \includegraphics[width=0.85\columnwidth]{images/random_forest_fea
     ture_ranking}
314: \caption{Random Forest Feature Ranking}
     \label{fig:random_feature_ranking}
403: \begin{figure}[H]
405: \includegraphics[width=0.8\columnwidth]{images/ridge_feature_rank
     ing_pos}
406: \caption{Ridge Algorithm - Top 10 Positive Features}
     \label{fig:ridge_feature_ranking_pos}
409: \begin{figure}[H]
411: \includegraphics[width=0.8\columnwidth]{images/ridge_feature_rank
```

```
     ing_neg}
412: \caption{Ridge Algorithm - Top 10 Negative Features}
     \label{fig:ridge_feature_ranking_neg}
419: \begin{table}[H]
422: \label{tab:xgb_param}
485: \begin{figure}[H]
487: \includegraphics[width=0.75\columnwidth]{images/xgboost_feature_i
     mportance}
488: \caption{XGBoost Feature Importance} \label{fig:xgb_feature_imp}
498: \begin{table}[H]
500: \label{tab:kaggle}

figures 14
tables 2
includegraphics 14
labels 14
refs 0
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 refered to: (refs >= labels)

Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors
```

```
59: Aarshay_Jain2016: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty year in 1
(There was 1 warning)

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

IoT and Big Data Analytics for Equipment Predictive Health Management (PHM)

Ashok Reddy Singam
Indiana University
711 N Park Ave
Bloomington, Indiana 47408
asingam@iu.edu

Anil Ravi
Indiana University
711 N Park Ave
Bloomington, Indiana 47408
anilravi@iu.edu

ABSTRACT

The predictive health management (PHM) is an enabling discipline consisting of technologies and methods to assess the reliability of a product in its actual life cycle conditions to determine the advent of failure and mitigate system risk. The PHM system will monitor environmental, operational, and performance related characteristics of the product and gathered data analyzed to assess product health and predict remaining life.

In this application, the industrial rotating equipment such as compressors, vacuum blowers, pumps, and valves etc. are considered to monitor and analyze their operational behavior. The product critical operational parameter data such as vibration, temperature, and load current will be collected from field sensors and analyzed to predict the failure using kNN machine learning classification algorithms. The data will be collected from the field using wireless sensors and stored on the cloud based AWS database server. The product data will be analyzed and made available to all stake holders to take appropriate preventive actions via web/mobile applications.

KEYWORDS

i523, HID333, HID337, KNN, IoT, Big Data, Analytics

1 INTRODUCTION

The PHM technology can be put within a broader business context by relating it to the Product-Service System (PSS) business model. PSS can be defined as an integrated combination of products and services where the emphasis is put on the 'sale of use 'rather than the 'sale of product '. Central to this new business model is a shift from selling a product, and its related spare parts as required, to selling a solution that supports customer needs in the form of a service delivering a fully maintained and useable product [2]. As shown in Figure 1, There are several wireless technologies such as 802.11, cellular, and short distance wireless protocols can be used to collect and send data to the centralized servers.Also, data can be stored in cloud based technologies such as AWS, Microsoft Azure, IBM and Google etc. for processing.

[Figure 1 about here.]

Problem Statement : in the manufacturing operations, automotive and other process industries rotating equipment such as pumps, valves, compressors, and blowers are commonly used equipment for various purposes. These equipment are severely suffered from wear and tear, bearing degradation, shaft misalignment, corrosion, and other mechanical breakdowns. Due to the limitations of wireless enabled sensors based data acquisition it was very difficult to collect this data in the past. Also, due to real-time nature of the data acquisition, it was a huge challenge to store the data locally and process the information for applying machine learning algorithms. All these technological and infrastructure limitations caused industrial equipment health monitoring had become one of the sector businesses are losing the money due to operation shutdowns and unplanned maintenance etc.

Solution Approach : with the wireless sensors and cloud based server technologies, it has become possible to deploy hundreds of sensors in the manufacturing plant and collect the data and store with minimal costs. Once the data is stored on the servers with high computing power, machine learning algorithms can be used to process the sensor data to predict the equipment failures with reasonable accuracy. This approach has been named as predictive or prognostics health management of the equipment which is widely available in the recent times due to the availability of technological infrastructure.

The PHM generally combines sensing, collecting, storing and analyzing of environmental, operational, and performance related parameters to assess the health of a product and predict remaining useful life. Assessing the health of a product provides information that can be used to meet several critical goals: (1) providing advance warning of failures; (2) minimizing unscheduled maintenance, extending maintenance cycles, and maintaining effectiveness through timely repair actions; (3) reducing the life cycle cost of equipment by decreasing inspection costs, downtime, and inventory; and (4) improving qualification and assisting in the design and logistical support of fielded and future systems [1].

The PHM is not a new concept, however, with the advent of sensors, machine learning algorithms, and computing capacity of the servers it has become more prevalent in the recent days. In this application, an attempt has been made to prove the concept of simple PHM implementation and use in real world applications. The application can be re-architected to address more complex products/systems with considerations of scalability, performance, cost and reliability. The limitations of the current application are described in the end of this report.

The parameter monitoring and the analysis of acquired data using prognostic models are fundamental steps for the PHM methods. The sensors are the essential devices used to monitor parameters and obtain long-term accurate information to provide anomaly detection, fault isolation, and rapid failure prediction [1].

Firstly, PHM requires monitoring a large number of product parameters to evaluate the health of a product. Depending on the complexity of the monitored product, it is possible to monitor thousands of parameters in the entire life cycle of the product to provide

the information required by PHM. These parameters include operational and environmental loads as well as the performance conditions of the product, for example, temperature, vibration, shock, pressure, acoustic levels, strain, stress, voltage, current, humidity levels, contaminant concentration, usage frequency, usage severity, usage time, power, and heat dissipation. In each case, a variety of monitoring features such as magnitude, variation, peak level, and rate of change may be required in order to obtain characteristics of parameters.

In this application, commonly used equipment in industrial and automobile operations such as air compressors, vacuum blowers, and smart valves are considered for analysis. The critical operational parameters of these products will be collected using applicable sensors from the field and fed to a database at regular intervals.

In general design, the frequency of data collection and storage depends on the number of parameters to be analyzed, cost of the system and operational behavior of the equipment. For this application, since products with rotating parts are considered, the critical parameters that would define the health of the equipment are: input or load current, internal ambient temperature, and vibration of the equipment.

The PHM application design process is shown in Figure 2, which describes various steps of the processes involved. For the implementation of this project, the sensor generated data is simulated using SQL scripts due to development time constraints. However, a detailed step-by-step approach is provided if we need to plug-in the sensor modules in to the application.

[Figure 2 about here.]

Data Acquisition Stage: It is required to have a description of a machine behaving normally that can be used for early detection of anomalies. This calls for a proper characterization of machine health. As part of this process, various methods are identified to extract health information from vibration measurements and investigate strengths and weaknesses of these methods as health descriptors. This stage will be the core part of PHM application where vibration data were experimentally obtained from a compressor using triaxial accelerometer to collect transverse, longitudinal and vertical axes vibration signals. For the experimental data collection, ACC301A triaxial accelerometer and National Instruments data acquisition system was used. A total of 8 parameters (1) input current (2) input voltage (3) internal ambient temperature (4) external ambient temperature (5) transverse vibration (6) longitudinal vibration (7) vertical axis vibration and (8) acquisition time were captured at 1 second rate, which generated about 65000 records. This data has been analyzed for identifying the feature classification.

Pre Processing stage: during this stage collected data will be filtered and processed for accuracy in order to adapt them to subsequent feature extraction stage. In this application, all the preprocessing has been done manually to validate the accuracy of the data based on the system conditions. Since spectral analysis of vibration signals are not done (one of the limitations for this application, captured in the end), the data generated from the compressor is considered as the primary frequency of the equipment (which is isolated from the rest of the attachments).

Feature extraction and selection stage: during this stage domain specific vibration spectral analysis has been performed but only considered time-domain behavior for various system operational conditions such as increased load, modified input voltage, and modified external ambient temperature etc. Based on the response of the machine vibration to various external conditions were noted down. This data is used to identify the following feature vectors.

- NORMAL OPERATION AT 30 DEG CENTIGRADE
- OVER CURRENT FAULT OPERATION
- OVER TEMPERATURE FAULT OPERATION
- INPUT OVER VOLTAGE FAULT OPERATION
- ABNORMAL OPERATION AT 30 DEG CENTIGRADE
- BEARING DEGRADATION OPERATION

kNN classification stage: this stage is core part of the PHM application, which will predict the unknown test data to be classified in to a known label based on the training data set using nearest neighbor algorithm.

Classifier performance evaluation stage: this stage will be used to evaluate the classifier accuracy of prediction. In this application, k-fold cross-validation method has been used to perform the evaluation.

The data is generated and made available in Oracle database on AWS cloud to perform analysis. The application developed in this project will consist of the following components:

- Sensor Data Generator
- Machine Learning Algorithm
- Big Data and IoT
- PHM Dashboard
- Decision Alerts
- Application Script

The following sections will describe the architectural and design aspects of the PHM system implementation in detail.

2 PROGNOSTICS MODEL EVALUATION

[3] The prediction is typically performed only after the *health* of the component or system deteriorates beyond a certain threshold. In this application, faults and failures are identified in the training data set. The faults identified are: Over current fault, over temperature fault and over voltage fault. If over current fault is occurred, the equipment will tend to draw higher current than nominal values which if continued further several times eventually leads to a permanent failure of the equipment. In this application, when motor bearing starts degrading, the first observation will be over current followed by over temperature conditions. Often times, that threshold is tripped because a fault occurs. A fault is a state of a component or system that deviates from the normal state such that the integrity of the component is outside of its required specification. A fault does not necessarily imply that the overall system does not operate anymore; however, the damage that characterizes the fault often grows under the influence of operations to a failure. The latter is the state at which the component or system does not meet its desired function anymore. It is the task of prognostics to estimate the time that it takes from the current time to the failed state, conditional on anticipated future usage. This would give operators access to information that has significant implications on system safety or cost of operations. Where safety is impacted, the ability to predict failure allows operators to take action that preserves the assets either through rescue operation or through remedial action

that avert failure altogether. Where minimizing cost of operations is the primary objective, predictive information allows operators to avert secondary damage, or to perform maintenance in the most cost-effective fashion. Often times, there is a mix of objectives that need to be optimized together, sometimes weighted by different preferences.

As emphasized above, predictive models evaluation needs to take domain specificities into account. Such specificities cover two aspects: capability of failure prediction and TTF estimation. From the point of view of TTF, it is desirable that a predictive model can generate alerts in a *targeted* time window prior to a failure. A model that predicts a failure too early leads to non-optimal component use [4] which will impact the reliability or availability of the system.

[Figure 3 about here.]

As shown in the Figure 3, the time to failure prediction will be estimated based on the classified result data set and alert the stakeholders to take relevant actions. The target alert zone will be identified based on the abnormal behavior of the equipment over the period.

3 APPLICATION DESIGN ANALYSIS

The PHM application in this project considered to use rotating equipment temperature, load current and vibration data for analyzing and predicting the future operational behavior. Vibration signals from rotating components are usually analyzed in the frequency domain, because significant peaks in the signal spectrum appear at frequencies that are related to the rotation frequency of the component. In this application, only time domain parameters with peak vibration magnitudes irrespective of the frequency component. The training data set consists of normal, abnormal, and fault conditions vibration patterns describes the system characteristics from which its status can be estimated. The PHM application for industrial equipment machine failure detection problem directly correlates to the pattern classification problem. From the vibration data collected, each accelerometer will output values of X, Y, and Z data then using a KNN we can similarly identify which vibration parameter(s) determines problems in our machines, or *likely to experience failure.* The typical defects or failures that can be detected are: machine imbalance, shaft misalignment, pumps cavitation, structural and rotating looseness, early stage bearing wear, gear teeth problems, and other high-frequency defects.

This application used *Sensor Data Gen* SQL script module to generate the sensor data and store in Oracle database on AWS. This is the critical module as we have not used the real data collection from the field. However, the sensor hardware and necessary environment to generate the data is identified and experimented to work with. A brief description about the hardware is provided in the end of this report.

The PHM application is designed such that the fundamental concepts can be verified to open a discussion on limitations, performance, scalability, ROI and reliability of the system.

The following sections describe the application design components with necessary implementation details:

3.1 Sensor Data Generator

The SQL data generator script is designed to generate training data as well test data for this application with following eleven parameters: Acquisition time, equipment name, part number, serial number, internal ambient temperature, external ambient temperature, input voltage, input current, and vibration data for x, y, and z axes. The following database design architecture followed for Sensor Data Gen module:

- Sensor Data Generator PL SQL Objects
  - Tables
    - SENSOR TRAIN DATA for storing training data
    - SENSOR TEST DATA for storing testing data
  - Views
    - SENSOR TRAIN DATA VIEW
    - SENSOR TEST DATA VIEW
  - Packages BIG DATA 503 PRJ PKG
    - Generate Test Set
    - Generate Train Set
    - Delete Data Set
    - Update Test Data Labels

3.2 Machine Learning Algorithm

3.2.1 Classifier evaluation. Typical classifier evaluation methods include ROC Curves, Reject Curves,Precision-Recall Curves, and Statistical Tests. The statistical tests consists of following methods to perform evaluation:

- Estimating the error rate of a classifier
- Comparing two classifiers
- Estimating the error rate of a learning algorithm
- Comparing two algorithms

Out of the listed statistical tests, the error rate estimation method is used in this application to evaluate the performance. An experimental data is used to estimate the error rate or accuracy of various classifiers. Then a comparison has been made to choose the classifier to use in the application.

The following list of performance for various classifiers is observed during the accuracy calculation. The same set of training data has been used for all the classifiers, which has resulted the following performance.

- LogisticRregression: 0.963636 (0.044536)
- KNN: 0.981818 (0.036364)
- DecisionTreeClassifier: 0.964394 (0.059656)
- SVM: 0.972727 (0.041660)

Based on the performance shown, kNN has been selected to use for this application.

[Figure 4 about here.]

3.2.2 k Nearest Neighbor - kNN. [5] In this application, neighbors-based classification is chosen to classify the unknown instance to the known trained labels. Neighbors-based classification does not attempt to construct a general internal model, but simply stores instances of the training data.

Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consisting of training observations (x,y) and would like to capture the relationship between x and y. More formally, our goal is to learn a function h:X -¿Y so that given an unseen observations x, h(x) can confidently predict the corresponding output y.

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given unseen observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by:

$$d(x, x') = \sqrt{(x1 - x1')^2 + (x2 - x2')^2 + ... + (xn - xn')^2}$$

But other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance. An alternate way of understanding KNN is by thinking about it as calculating a decision boundary (i.e. boundaries for more than 2 classes) which is then used to classify new points.

In the application design, sci-kit open source python libraries are used for implementing the kNN algorithms. Scikit is built on NumPy, SciPy, and matplotlib. The k-neighbors classification in KNeighborsClassifier is the more commonly used of the two techniques. The optimal choice of the value k is highly data-dependent: in general a larger k suppresses the effects of noise, but makes the classification boundaries less distinct.

The sklearn.neighbors.KNeighborsClassifier class has the following methods, which are used in the application deisgn:

- fit: Fit the model using X as training data and y as target values.
- get params: Fit the model using X as training data and y as target values.
- kneighbors: Finds the K-neighbors of a point.
- kneighbors graph: Computes the (weighted) graph of k-Neighbors for points in X.
- predict: Predict the class labels for the provided data.
- predict proba: Return probability estimates for the test data X.
- score: Returns the mean accuracy on the given test data and labels.
- set params: Set the parameters of this estimator.

3.2.3 K-fold cross-validation. To estimate the test error in the model, a cross-validation approach followed in which a subset of the training set will be holding out from the fitting process. This subset, called the validation set, can be used to select the appropriate level of flexibility of our algorithm. There are different validation approaches that are used in practice, and we will be exploring one of the more popular ones called **k-fold cross validation.** The k-fold cross validation (the k is totally unrelated to K) involves randomly dividing the training set into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds. The misclassification rate is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error which are then averaged out.

In this application, an average k-fold cross validation accuracy of 0.99 achieved, which is explained in the appendix section of the report.

3.3 Big Data and IoT

In PHM systems big data is characterized by one or more 3Vs: volume, velocity and variety due to streaming of real-time IoT sensors. Most of the IoT systems present challenges in combinations of velocity and volume. The important feature of the IoT application is that by observing the behavior of "many things"it will be possible to gain important insights, optimize processes, etc. This requires storing all the events (velocity and volume challenge) to run analytical queries over the stored events and perform analytics (data mining and machine learning) over the data to gain insights. In general PHM applications, data will be collected through field sensors at specific rate which accounts for large amount of data per day in the order of multi-million records. This data will be stored in any NOSQL or RDBMS based database for storage and processing. Since the big data infrastructure is much reliable and available widely from multiple vendors, it would help to build complex PHM systems with large number of feature vectors for classification.

In this application, for the demonstration of the concept, the real vibration data from the compressor equipment has been collected via accelerometer sensors. This vibration data has been analyzed in time domain and established the labels based on the compressor design performance parameters. Later, this data analysis is used to design a SQL script for generating training and test data sets. However, the real-time PHM system will have continuous streaming of data coming from hundreds of devices at faster rates (in the order of milliseconds to tens of seconds). This data needs to be captured by reliable and scalable platforms such as AWS IoT or similar and use the machine learning algorithms to classify the unknown data.

3.4 PHM Dashboard

Once all the test data set has been classified in to appropriate labels, the prediction of the failure can be performed based on the trending of the equipment behavior over the period. In order to understand the equipment performance insight, following queries will be used on the classified data:

- Faults Reported by Equipment Part Number
- Faults Reported by Serial Number
- Abnormal Behavior by Equipment Part Number
- Abnormal Behavior by Serial Number over the period range

There can be more application specific information obtained from classified data set to take various decisions. Figure 4 shows various PHM data analytics for this project. Figure 4(a) displays all the serial numbers of equipment 1 with bearing degradation problems. The X axis gives serial numbers while the Y axis gives number of occurrences of bearing degradation for that particular serial number. Similarly Figure 4 (b) and 4 (c) give the details of over temperature and over current faults of various serial numbers.

As part of data visualization, result data file is queried based on the analytics metrics interested. The python matplotlib package has been used to draw the charts as needed for showing the analytics. In real world application, a more sophisticated business intelligence tools such as Tableau, Microsoft BI, and Amazon Quick Sight can be

used to show the PHM dashboards. These dashboards are targeted for business users so that they will be able to customize the views, add filters and drill down in to specific information as needed.

[Figure 5 about here.]

3.5 Decision Alerts

Once the results data set has been generated by the prediction algorithm, and then based on the analytics queries, PHM system can send out the alert messages to appropriate stakeholders. The typical messages include the following s minimum:

- SN 10002: Faulted X times on over temperature in last Y days, needs maintenance to clean the filter
- SN 10005: Consistently indicating bearing degradation from last X days, needs lubrication maintenance
- SN 10009: Consistently drawing over current from last X days, needs mechanical load maintenance

In this application all the unseen test data is classified and labeled in the result data file. However, in real world application, along with the dashboards a comprehensive alerting capabilities can be built. The application checks for out of range alert conditions on selected incoming report parameters, looking for warning or alarm conditions that are higher or lower than expected under normal operating conditions.

3.6 Application Script

All the application specific python modules are captured in a ipynb file which can be run from Jupyter notebook or Github. The location of the file is described in appendix b. This application specific data has been created/generated on AWS cloud database, which will be accessed during the run time by python notebook.

4 APPLICATION LIMITATIONS

The PHM application developed in this project has several limitations. Typically, PHM applications suffer from the prediction accuracy rate which influences the ability to take decisions that will have broader impact on the business operations and financial aspects. However, with the advanced machine learning classifiers and model evaluation methods this can be addressed to achieve reasonable confidence. Following are some of the limitations of this application, which can be addressed and improved in large real-time PHM systems.

Data acquisition hardware : in this application the data is not collected from real-time sensors for voltage, current, temperature and vibration data. There will be inherent accuracy in the raw data generated by SQL script. However, a sample vibration dataset has been collected from the field, which is used as basis to generate the simulated data set.

Feature extraction analysis : the equipment performance parameters of interest need to be down selected from large set of incoming parameter data.

When analyzing vibration data in the time domain only few parameters are available in quantifying the strength of a vibration profile: amplitude, peak-to-peak value, and RMS. The amplitude is valuable for shock events but it does not take into account the time duration and thus the energy in the event. The same is true for peak-to-peak with the added benefit of providing the maximum excursion of the wave, useful when looking at displacement information, specifically clearances. The RMS value is generally the most useful because it is directly related to the energy content of the vibration profile and thus the destructive capability of the vibration.

This requires in-depth domain specific analysis, in this case a detailed mathematical modeling of vibration spectral analysis to precisely select the features and corresponding behavior patterns. Such analytical data should be used for training data feature set. In this application, a primitive approach of time-domain analysis of vibration magnitudes used for determining features. However, in real application these features need to be mathematically analyzed to identify the features that represent the system behavior as close as possible.

Model accuracy and scoring : the kNN algorithm used in this application validated using k-cross fold cross-validation. There are several other model evaluation and scoring methods such as accuracy (or error rate), True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), sensitivity etc. These metrics provide a simple and effective way to measure the performance of a classifier. This application can be further improved by applying more performance measurement methods to increase the effectiveness of the algorithm design.

Scalability : the application designed in this project is very primitive to understand the basic concepts of PHM and kNN classifier implementation. This application cannot be used for PHM application in business use. To implement real world PHM application, a more comprehensive design needed by considering modularity, service oriented architecture, large number of sensors integration, big data and analytics integration etc.

5 RECOMMENDATIONS

Generality : since each rotating equipment vibrates in a different manner, a monitoring method needs to be retrained for each machine. The training on several repeated measurements on several similar equipment in several operating modes may allow for a more general monitoring method.

Feature extraction and dimensionality : In this application it has been assumed that a proper feature (selection) has been chosen, such that the feature dimensionality is not too high. If the data lies in a subspace, application of an initial dimensionality reduction may be a good idea. It is highly recommended to perform spectral analysis on vibration data and identify various fault frequencies and their sources. This would help to extract the optimized feature vectors for the given application followed by selecting the more relevant ones.

Model evaluation : classifier accuracy and effectiveness will be varied based on the test data set. It is highly recommended to evaluate multiple models with appropriate test data to choose the best classifier for the given application.

Domain specific modeling : It is highly recommended to perform more and more domain-oriented feature vector analysis to meet the needs of predictive model evaluation for PHM applications. Domain-oriented approaches helpful and useful in evaluating classifier for applications. Generic evaluation methods could help developers in investigating overall performance of a model from the statistical viewpoint at the initial stage of model development.

Domain-oriented approaches should be further used to evaluate the usefulness and business value [4].

6 CONCLUSION

In this project, the problem statement around industrial rotating equipment maintenance is described and solution principle to address the same using PHM concept is defined, experimented and results are discussed. Since this application is developed to prove the only concept but not the complete solution a section with limitations and recommendations for real world system development is described. Overall, PHM application with kNN classifier algorithm and cross validation accuracy of 0.99 has been implemented, verified and results are analyzed for business decisions.

ACKNOWLEDGMENTS

The authors would like to thank professor Gregor von Laszewski and his team for providing *LaTex* templates and assistance with the *JabRef* tool to organize references.

REFERENCES

[1] Shunfeng Cheng, Michael H. Azarian, and Michael G. Pecht. 2010. Sensor Systems for Prognostics and Health Management. (2010), 24 pages. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3247731/pdf/sensors-10-05774.pdf

[2] Tonci Grubic, Ian Jennions, and Tim Baines. 2009. The Interaction of PSS and PHM - a mutual benefit case. (2009), 10 pages. https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2009/phmc_09_49.pdf

[3] Abhinav Saxena, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. 2009. Sensor Systems for Prognostics and Health Management. (2009), 16 pages. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100023445.pdf

[4] Chunsheng Yang, Yanni Zou, Jie Liu, and Kyle R Mulligan. 2014. Predictive Model Evaluation for PHM. (2014), 11 pages. https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2014/ijphm_14_019.pdf

[5] Kevin Zakka. 2016. A Complete Guide to K-Nearest-Neighbors with Applications in Python and R. (2016). https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/

A WORK BREAKDOWN

A.1 HID 333:Anil Ravi

- Identified Project topic.
- Created architecture of the application.
- Ran experimental test to collect vibration data
- Extracted and analyzed feature vectors
- Studied, designed and reviewed kNN algorithm
- Created draft project report
- Reviewed the draft project report.

A.2 HID 337:Ashok Reddy Singam

- Implemented sensor data generation SQL script.
- Implemented kNN algorithm in Python
- Implemented k-fold cross validation design
- Created data analytics charts
- Reviewed the draft project report.

B CODE REFERENCE

All code, notebooks and files for this project can be found in the githup repository: https://github.com/bigdata-i523/hid337/blob/master/project/jupyter

List of Figures

1 System Architecture 8
2 PHM Design Process 8
3 Time relation between alert time and failure time 9
4 Algorithm performance 9
5 PHM Data Analytics 10

Figure 1: System Architecture

Figure 2: PHM Design Process

Figure 3: Time relation between alert time and failure time

Figure 4: Algorithm performance

9

(a) Bearing Degradation

(b) Over Temperature Fault

```
bibtext report
===========================================

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

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


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


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


latex report
===========================================

        [2017-12-04 22.02.41] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
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 289.10326pt.
Typesetting of "report.tex" completed in 1.0s.

================================================================================
Compliance Report
================================================================================


name: Ashok Reddy Singam
hid:  337
paper1: Nov 01 17 100%
paper2: Nov 06 17 100%
project: Dec 04 17 100%

yamlcheck
--------------------------------------------------------------------------------
```

```
wordcount
--------------------------------------------------------------------------------

10
wc 337 project 10 4977 report.tex
wc 337 project 10 4893 report.pdf
wc 337 project 10 226 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
--------------------------------------------------------------------------------

38: \begin{figure}
39: \includegraphics[width=1.0\columnwidth]{images/system_architecture
    }
40: \caption{System Architecture} \label{fig:Figure1}
64: \begin{figure}
65: \includegraphics[width=1.0\columnwidth]{images/phm_process_1}
66: \caption{PHM Design Process} \label{fig:Figure2}
109: \begin{figure}
110: \includegraphics[width=1.0\columnwidth]{images/ttf_1}
111: \caption{Time relation between alert time and failure time}
     \label{fig:Figure3}
174: \begin{figure}
175: \includegraphics[width=0.9\columnwidth]{images/algperformance}
```

```
176: \caption{Algorithm performance} \label{fig:Figure5}
238: \begin{figure}
241: \includegraphics[width=1.0\columnwidth]{images/DEGRADE}
242: \caption{Bearing Degradation} \label{sfig:sfig4a}
246: \includegraphics[width=1.0\columnwidth]{images/OVERTEMP}
247: \caption{Over Temperature Fault} \label{sfig:sfig4b}
251: \includegraphics[width=1.0\columnwidth]{images/OVERCURR}
252: \caption{Over current Fault} \label{sfig:sfig4c}
255: \caption{PHM Data Analytics} \label{fig:Figure4}

figures 5
tables 0
includegraphics 7
labels 8
refs 0
floats 5

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 refered to: (refs >= labels)

Label/ref check
34: The PHM technology can be put within a broader business context by
    relating it to the Product-Service System (PSS) business model.
    PSS can be defined as an integrated combination of products and
    services where the emphasis is put on the \lq sale of use \rq
    rather than the \lq sale of product \rq. Central to this new
    business model is a shift from selling a product, and its related
    spare parts as required, to selling a solution that supports
    customer needs in the form of a service delivering a fully
    maintained and useable product \cite{Tonci2009}. As shown in
    Figure 1, There are several wireless technologies such as 802.11,
    cellular, and short distance wireless protocols  can be used to
    collect and send data to the centralized servers.Also, data can be
    stored in cloud based technologies such as AWS, Microsoft Azure,
    IBM and Google etc. for processing.
59: The PHM application design process is shown in Figure 2, which
    describes various steps of the processes involved. For the
    implementation of this project, the sensor generated data is
    simulated using SQL scripts due to development time constraints.
    However, a detailed step-by-step approach is provided if we need
    to plug-in the sensor modules in to the application.
113: As shown in the Figure 3, the time to failure prediction will be
     estimated based on the classified result data set and alert the
     stakeholders to take relevant actions. The target alert zone will
```

```
     be identified based on the abnormal behavior of the equipment
     over the period.
233: There can be more application specific information obtained from
     classified data set to take various decisions. Figure 4 shows
     various PHM data analytics for this project. Figure 4(a) displays
     all the serial numbers of equipment 1 with bearing degradation
     problems. The X axis gives serial numbers while the Y axis gives
     number of occurrences of bearing degradation for that particular
     serial number. Similarly Figure 4 (b) and 4 (c) give the details
     of over temperature and over current faults of various serial
     numbers.
passed: False -> labels or refs used wrong

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



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 comunication between these devices, and to analyse the data collected by these devices, which is made possible by the use of a scalable lightweight communicatoin protocol such as MQTT, which is easy to use, data agnostic, and application independent. We look at one such appllication 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 trafitional HTTP requests efficiently, and as the number of devices increases, sending an HTTP request to each of the dvices in order to get data may not be efficient [3][10].

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 ad available memory are limited [38][21]. The Eclipse Paho Project curently provides suport 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 recieves over mqtt, and drives the car accordigly. another program running on the raspberry pi uses the rasperry pi onboard camera to capture pictures and send them back to the desktop to helo 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 conect to the MQTT broker.

We can use multiple such cars and controlling devices to control the cars independently or from a commn device to drive mltiple 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.

[27] 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 [13]

Zheng Wang used raspberry pi in [37] 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 predigt 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 [15].

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 vatious home automation systema and provides an application interface to control those devices [24]. It allows integration of various devices with MQTT.

The FloodNet project at University of Southampton [11] aims at "providing a pervasive, continuous embedded monitoring presence". The system is intelligent and obtains " environmental self-awreness 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 [31] 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 [28].

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 [18].

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 [18]. 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 [38].

MQTT allows the publisher and subscriber to respond to messages withthe 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) [19][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 [19] [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 [19] [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 is 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" [26] [20]. 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 [20].

Transport Layer Security (TLS) and Secure Sockets Layer (SSL) are cryptographic protocols that establish a 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 [20]. "Most MQTT brokers provide an option to use TLS instead of plain TCP and port 8883 has been standardized for secured MQTT connections" [26].

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 [26][20].

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 ans the user and are signed by a trusted authorization server [9].

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 [9]. 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 [35] to get real time data streams and allow analytics and online machine learning in a fault tolerant manner [41].

ELK stack (elastic-search, logstash and kibana) is an opensource 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 [33] and [32].

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 [8].

3.2 Raspberry Pi

The raspberry pi is a credit card sized development board that was developed by Eben Upton with the goal to crate a low cost device that can be used for education and prototypiing [25]. Since its creation the board has been adapted for various different projects by educators hobbyists and in the industry [30]. 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 connnecting 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 [30].

The raspberry pi comes in various models, Model A+, which is one of the smallest form factors, rasbperry 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 [17][16].

The raspberry pi Zero is the deveopment board that has the smallest form factor. Even though the the raspberry pi zero includes no ehternet 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 [29].

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 cantered 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 [25].

Variuos 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 alows simple interface with may 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 otor 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 [40].

Stepper motors can be controlled in various ways, dependng on the application. Figure 1 shows how a steppper 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 [14].

[Figure 1 about here.]

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 adjescent 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 [36].

[Figure 2 about here.]

With full stepping however, the transition between two consecutive steps is not very smooth. Therefore, a technique called Half stepping is used, where two adjescent 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 [14].

[Figure 3 about here.]

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 steper 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 [12][34].

3.4 Opencv

The Open Source Computer Vision library (openCV) is a library of functions aimed at real time computer vision and macine learning and providing a common infrastructure to allow fast progress in the field of computer vision and machine perception [39][22].

The library was originally build 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 [22]. for Python the library has been built on top of Numpy, a library that optimizes matrix and vector operations, and takes advantade of MMX and SSE instructions whenever possible. For C++ the library uses the Standart 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 extention to the library built collaboratively by the community contains advanced algorithms that allow processing video in real time [22].

OpenCV is widely used in the industry by startups as well as well established organizations like Google, Yhoo, Microdoft, IBM and Intel [22].

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 [23][42]. 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 accordig to he message it recieves from the desktop.

There are two programs running on pi, controller_stepper_sub.py ans 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 subsciber controller_stepper_sub.py runningon 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 ocnverted to greyscale, and opencv is used to perform face detection using Haar Cascades. If a face is found, a box is dran 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 fo the navigation of the robot car remotely figure 4.

[Figure 4 about here.]

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 contunuous 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

5 RESULTS

This section covers the setup instructions for the project and the observations.

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 prvided 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 raspberr pi GPIO pins , and assumes that for the left motor, the pins IN1, IN2, IN3, IN4 are conected 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 5

  [Figure 5 about here.]

- 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 an 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 ins 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 beween the raspberry pi and the desktop controller application is pretty seemless. 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 in 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 mush a motor should turn when it recieves a message. If the motor is not allowed to turn long enough, then between two mesages the motor will be idle and if it is turned longe than the interval between two messages, there can be conflicts if in response to each of the messages the subscriber runing on the raspberry pi tries to set a different step on the motor. Therefore, the movement can seem a litle 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 colflicts 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 commmunication

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 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 regognised only if they are resent inthe image in the same orientation as that in the trainig 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 ans rotations. A better option would be to Concolutional 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 diffrent 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 deveopment 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 wireles 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 providign the ncessay ardwawre to complete the project.

The author would also like to thank the associate istructors of th 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. (augist 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
[5] eclipse. [n. d.]. mqtt broker. eclipse mosquitto website. ([n. d.]). https://mosquitto.org/
[6] eclipse paho. [n. d.]. Python Client - documentation. eclipse paho wensite. ([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] erlang mqtt. [n. d.]. erlang mqtt broker. wmqtt website. ([n. d.]). http://emqtt.io/docs/v2/index.html
[9] 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
[10] hivemq. [n. d.]. intrwebsite mqtt. hivemq website. ([n. d.]). https://www.hivemq.com/blog/mqtt-essentials-part-1-introducing-mqtt
[11] University of Southampton IAM group. 2005. FloodNet. IAM group website. (April 2005). http://www.iam.ecs.soton.ac.uk/projects/297.html
[12] Elegoo Industeies. 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/
[13] Dexter Industries. 2017. GoPiGo Build and Program Your Own Robot. dexter industries website. (2017). https://www.dexterindustries.com/gopigo3/
[14] Images Scientific Instrumentation. 2017. How Stepper Motors Work. imagesco.com website. (2017). http://www.imagesco.com/articles/picstepper/02.html
[15] 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.
[16] 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
[17] makershed.com. 2016. Raspberry pi comparison chart. makershed.com website. (2016). https://www.makershed.com/pages/raspberry-pi-comparison-chart
[18] Hive mq. [n. d.]. MQTT Essentials Part 2: Publish & Subscribe. HiveMQ website. ([n. d.]). https://www.hivemq.com/blog/mqtt-essentials-part2-publish-subscribe
[19] 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
[20] Hive Mq. [n. d.]. MQTT Security Fundamentals: TLS / SSL. hive mq website. ([n. d.]). https://www.hivemq.com/blog/mqtt-security-fundamentals-tls-ssl
[21] Mqtt. [n. d.]. Mqtt official website. mqtt official website. ([n. d.]). http://mqtt.org/
[22] Opencv. 2017. About. opencv.org website. (2017). https://opencv.org/about.html
[23] 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
[24] OpenHab. 2017. What is openHAB? openhab website. (November 2017). https://www.openhab.org/introduction.html
[25] opensource.com. 2015. What is a Raspberry Pi. opensource.com website. (March 2015). https://opensource.com/resources/raspberry-pi
[26] 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/
[27] pythonprogramming.net. 2014. Robotics with Python Raspberry Pi and GoPiGo Introduction. pythonprogramming.net. (April 2014). https://pythonprogramming.net/robotics-raspberry-pi-tutorial-gopigo-introduction/
[28] 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/
[29] raspberrypi.org. 2015. Raspberry Pi Zero: the 5 dollar computer. raspberrypi.org. (November 2015). https://www.raspberrypi.org/blog/raspberry-pi-zero/
[30] raspberrypi.org. 2015. What is a Raspberry pi. raspberrypi.org website. (May 2015). https://www.raspberrypi.org/help/what-%20is-a-raspberry-pi/
[31] 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
[32] smart factory. 2016. MQTT and Kibana fi?! 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/
[33] smart factory. 2016. Storing IoT data using open source. MQTT and ElasticSearch fi?! Tutorial. smart factory website. (october 2016). https://smart-factory.net/mqtt-elasticsearch-setup/
[34] 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/
[35] Apache storm. [n. d.]. Storm MQTT Integration. Apache storm website. ([n. d.]). http://storm.apache.org/releases/1.1.0/storm-mqtt.html
[36] Built to spec. 2015. Understanding Stepper Motors Part I fi?! 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/
[37] Zheng Wang. 2015. Self Driving RC Car. Zheng Wang wordpress website. (August 2015). https://zhengludwig.wordpress.com/projects/self-driving-rc-car/
[38] 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].
[39] Wikipedia. 2017. OpenCV — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=OpenCV&oldid=811519079 [Online; accessed 4-December-2017].
[40] Wikipedia. 2017. Stepper motor — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=Stepper_motor&oldid=811220740 [Online; accessed 4-December-2017].
[41] Wikipedia. 2017. Storm (event processor) — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=Storm_(event_processor)&oldid=808771136 [Online; accessed 6-November-2017].
[42] Wikipedia. 2017. Violafi?!Jones object detection framework — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=Viola%E2%80%93Jones_object_detection_framework&oldid=808683512 [Online; accessed 4-December-2017].

A ISSUES

DONE:

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

A.1 Assignment Submission Issues

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

A.2 Uncaught Bibliography Errors

Missing bibliography file generated by JabRef

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

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

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

Other formatting issues

A.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

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

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

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

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

Use a space after . , :

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

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

A.5 Citation Issues and Plagiarism

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

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

A.6 Character Errors

Erroneous use of quotation marks, i.e. use "quotes" , instead of " "

To emphasize a word, use *emphasize* and not "quote"

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

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.

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

A.7 Structural Issues

Acknowledgement section missing

Incorrect README file

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

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

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

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

A.8 Details about the Figures and Tables

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

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

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

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

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

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

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

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

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

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

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

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

Do not use textwidth as a parameter for includegraphics

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

e.g.

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

re

List of Figures

1 Working of a steper motor : Full stepping using one coil at a time [14] 9
2 Working of a steper motor : Full stepping using two coils at a time [36] 10
3 Working of a steper motor : Half stepping [14] 11
4 Architecture of the Application 12
5 Connection Diagram 13

Figure 1: Working of a steper motor : Full stepping using one coil at a time [14]

Figure 2: Working of a steper motor : Full stepping using two coils at a time [36]

Figure 3: Working of a steper motor : Half stepping [14]

Figure 4: Architecture of the Application

Figure 5: Connection Diagram

```
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
===========================================

        [2017-12-04 22.00.22] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
```

```
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.3s.
../README.yml
  9:81      error    line too long (83 > 80 characters)  (line-length)
  10:81     error    line too long (81 > 80 characters)  (line-length)
  11:81     error    line too long (82 > 80 characters)  (line-length)
  12:81     error    line too long (81 > 80 characters)  (line-length)
  13:81     error    line too long (81 > 80 characters)  (line-length)
  27:81     error    line too long (81 > 80 characters)  (line-length)
  27:81     error    trailing spaces  (trailing-spaces)
  28:81     error    line too long (83 > 80 characters)  (line-length)
  29:80     error    trailing spaces  (trailing-spaces)
  30:81     error    line too long (83 > 80 characters)  (line-length)
  30:83     error    trailing spaces  (trailing-spaces)
  31:81     error    line too long (83 > 80 characters)  (line-length)
  31:83     error    trailing spaces  (trailing-spaces)
  32:81     error    line too long (89 > 80 characters)  (line-length)
  33:81     error    line too long (89 > 80 characters)  (line-length)
  34:81     error    line too long (89 > 80 characters)  (line-length)
  34:89     error    trailing spaces  (trailing-spaces)
  47:81     error    line too long (83 > 80 characters)  (line-length)
  47:83     error    trailing spaces  (trailing-spaces)
  49:81     error    line too long (83 > 80 characters)  (line-length)
  49:83     error    trailing spaces  (trailing-spaces)
  50:81     error    line too long (86 > 80 characters)  (line-length)
  50:86     error    trailing spaces  (trailing-spaces)
  52:81     error    line too long (90 > 80 characters)  (line-length)
  53:81     error    line too long (92 > 80 characters)  (line-length)
  63:1      error    trailing spaces  (trailing-spaces)


================================================================================
Compliance Report
================================================================================


name: Arnav, Arnav
hid:  201
paper1: 20th Oct 2017 100%
```

```
paper2: 80%
project: 100%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

13
wc 201 project 13 4845 report.tex
wc 201 project 13 6009 report.pdf
wc 201 project 13 1780 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
--------------------------------------------------------------------------------

228: Figure \ref{f:stepper1} shows how a steppper 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.pd
     f}
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 adjescent 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.pd
     f}
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 adjescent 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.pd
     f}
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 ocnverted to greyscale, and opencv is used to perform
     face detection using Haar Cascades. If a face is found, a box is
     dran 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 fo 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}
297: \ref{f:connection}
299: \begin{figure}[!ht]
```

```
300: \centering\includegraphics[width=\columnwidth]{images/connection.
     pdf}
301: \caption{Connection Diagram}\label{f:connection}

figures 5
tables 0
includegraphics 5
labels 5
refs 5
floats 5

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

Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



WARNING: 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 adjescent 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



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 "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
```

```
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3131 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
```

```
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 44 error messages)
make[2]: *** [bibtex] Error 2

latex report
============================================

        [2017-12-04 13.49.02] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 1.4s.
../README.yml
  37:81     error    line too long (85 > 80 characters)  (line-length)
  37:85     error    trailing spaces  (trailing-spaces)
  38:81     error    line too long (85 > 80 characters)  (line-length)
  38:85     error    trailing spaces  (trailing-spaces)
  39:81     error    line too long (87 > 80 characters)  (line-length)
  40:81     error    line too long (89 > 80 characters)  (line-length)
```

```
  41:81     error    line too long (89 > 80 characters)  (line-length)
  41:89     error    trailing spaces  (trailing-spaces)
  42:53     error    trailing spaces  (trailing-spaces)


================================================================================
Compliance Report
================================================================================


name: Syam Sundar Herle Parampali Sreenath
hid:  219
paper1: 100%
paper2: 100%
project: Dec 04 17 0%

yamlcheck
--------------------------------------------------------------------------------


wordcount
--------------------------------------------------------------------------------

10
wc 219 project 10 5929 report.tex
wc 219 project 10 7733 report.pdf
wc 219 project 10 3283 report.bib

find "
--------------------------------------------------------------------------------

161: Increasing access to Naloxone, an opioid antagonist that rapidly
     reverses the opioid overdose damage, may be the most important
     immediate solution to reducing opioid-related deaths
     \cite{Hawk01}. Between 1998 and 2014, 52,283 naloxone kits were
     distributed among the 30 states with naloxone distribution
     programs resulting in 26,453 overdose reversals \cite{Hawk01}. 27
     states have "third-party prescription" laws that allow physicians
     to prescribe Naloxone to family and friends of individuals with
     an opioid addiction \cite{Hawk01}. To further reduce opioid-
     related deaths states must reduce malpractice liability for
     physicians prescribing Naloxone and make Naloxone available
     without a prescription \cite{Hawk01}.

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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True
```

```
below_check
-------------------------------------------------------------------------------


bibtex
-------------------------------------------------------------------------------


label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
```

```
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3131 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3131 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3229 of file ACM-Reference-Format.bst
Name 2 in "Kathryn Hawk and Federico E Vaca, and Gail D'Onofrio" has a comma at the end
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Kakavand, H., Kost De Serves, N., Chilton, B" for entry pa
```

```
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Rudd RA, Seth P, David F, Scholl L" for entry opsis10
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
Name 1 in "Steve Cheng, Matthias Daub, Axel Domeyer, and Martin Lundqvist" has a comma a
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 44 error messages)

bibtex_empty_fields
--------------------------------------------------------------------------------


entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

15: note            = "",

42: note            = "",

76: note            = "",

436: pages   = "",

437: doi             = "",

439: note            = "",

passed: False

ascii
--------------------------------------------------------------------------------


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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

Big Data Analytics on Food Products Around the World

Karthik Vegi
Indiana University Bloomington
College Mall Apartments
Bloomington, Indiana 47401
kvegi@iu.com

Nisha Chandwani
Indiana University Bloomington
Park Doral Apartments
Bloomington, Indiana 47408
nchandwa@iu.edu

ABSTRACT

Food is one of the basic necessities of human-being. It helps us gain energy to recharge our body to do the daily activities like moving, playing, and thinking. From being a cave man to producing wide variety of foods, we have come a long way. The civilizations shaped the food habits of the world and there is a lot of variance in the food habits across countries. We analyze the *Open Food Facts* database that gathers information on food products from around the world to unearth some food habits of the world and we predict the food grade based on the nutrition facts of the food products.

KEYWORDS

i523, hid231, hid203, big data, food habits, food products, nutrition

1 INTRODUCTION

Open Food Facts is a non-profit initiative started by Stephane Gigandet and run by thousands of volunteers around the world. Any person around the world can contribute to the database by simply scanning a product using a mobile app which is made available to IOS and Android. This massive database of food products opens up a lot of opportunities to analyze the food products around the world and understand their food habits. We are particularly interested in the consumption of nutrients that come along with the food items across the world, the composition of different fat content, and the prediction of nutrition grade based on the nutrients.

2 FOOD ANALYSIS: IMPORTANCE AND RELATED WORK

In recent times, more and more companies try to market their food as low-fat or low-calories in order to fool consumers into buying their products. The increasing concern of public health has led to a significant interest in detecting the health-related properties of food products [2]. Thus, there is no question about the importance of analysis of the nutrition grade and food safety in today's world. The analysis of food requires more robust and efficient methodologies in order to ensure the quality and safety of the food products [2]. Previous methods based on the so-called wet-chemistry have now evolved into more powerful techniques which are used in the food laboratories These methods provide massive improvement in analytical accuracy thus expanding the limits of food applications [2]. The traditional methods of food analysis can be classified based on the underlying principle. Some of these categories are spectroscopic, biological, electrochemical, supercritical fluid chromatography [2]. All these techniques provide information about the sample under study and this information is derived from a specific physical-chemical interaction [2]. A different approach of analyzing and detecting the food quality is by using machine learning techniques. We will discuss one of these modern methods of food analysis which can be widely used across countries.

3 ANALYSIS OF FAT CONTENT IN FOOD

Fat is definitely a nutrient that the body needs and are essential to aid in cell growth, help with energy generation, maintaining body temperature, protect organs, help absorb other essential nutrients that aid in producing energy, improve blood cholesterol level, help reduce inflammation in case of injury, and help in storing energy that can be used for survival when you go without food for few days [1]. But, we do need to keep a track of the consumption because anything that is remotely excess leads to a variety of serious health issues [1].

3.1 Dietary Fats

There are different types of fat fi?! some are good and some are bad and some needs to be taken within a certain limit [1].

3.1.1 Saturated Fat. More intake of saturated fats results in the cholesterol levels in the blood which increases the risk of heart related diseases [1]. The American Heart Association suggests around 5 percent of daily calories from foods containing saturated fat [1]. Meat, cheese and milk are some of the sources of saturated fat [1].

3.1.2 Trans Fat. Any type of trans fat whether it is natural or artificial is not good [1]. The reason why food manufacturers use trans-fat is because they are less expensive, can be produced artificially, easy to use with other ingredients, last for a long time and also aid in improving the taste of the food [1]. Trans fats raise the bad fat levels and decreases the good fat levels [1]. The American Heart suggests to completely cut off trans-fat from the diet [1].

3.1.3 Monounsaturated Fat. Monounsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and there by decrease the risk of heart diseases [1]. They also help in gaining vitamin E which is a good nutrient that acts as antioxidant [1]. Olive oil, avocados, and sesame oil are some of the sources for monounsaturated fats [1].

3.1.4 Polyunsaturated Fat. Polyunsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and there by decrease the risk of heart diseases [1]. They also provide some nutrients that are essential for the body [1]. Soybean oil and sunflower oil are some of the sources for polyunsaturated fats [1].

3.2 Data Cleaning and Transformation

To make the analysis more interesting, the top 20 countries with most value counts for the attributes have been considered. The countries with names combined with other countries were also cleaned in the process. The data set was analyzed was missing values and the attributes with more than 60 percent missing values were removed from the analysis to add consistency. Only the columns that are meaningful in the analysis were retained and the rest were removed from further analysis.

We then display the top 5 countries as a pie-chart and the 5 countries are namely United States, France, Switzerland, Germany, and Spain as shown in Figure 1.

[Figure 1 about here.]

We then impute all the null values with zeroes and we then check the dietary fat content in the foods and check the top countries with fat content using a histogram. The analysis with respect to the fat countries is as follows

3.3 Data Analysis

The top 5 countries with most fat content in the food items are Serbia, United States, Switzerland, Germany, and Sweden as shown in Figure 2.

[Figure 2 about here.]

The top 5 countries with most saturated fat content in the food items are Serbia, United States, Germany, France and Switzerland as shown in Figure 3

[Figure 3 about here.]

The top 5 countries with most trans-fat content in the food items are United States, Brazil, Canada, Australia, Russia, and Serbia as shown in Figure 4

[Figure 4 about here.]

The top 5 countries with most cholesterol content in the food items are United States, Canada, Portugal, Brazil, France, and Italy as shown in Figure 5

[Figure 5 about here.]

4 ANALYSIS OF SUGAR AND SALT CONTENT

5 NUTRITION GRADE LABELLING SYSTEM

France recently took a decision to implement a nutri-score system which will use a color coding mechanism to label the food products that will help consumers know the nutrition grade of the product [3]. The World Health Organization regional office for Europe as a part of its 5 year action plan from 2015-2020 recommends a labelling mechanism for the consumers to know about the quality of the food products at a first glance [3]. This will not only make it easier for the consumers to pick healthier options but it will also regulate food manufacturers to resort to healthier ingredients instead of going for low cost artificial or less healthier ingredients [3].

France after United Kingdom became the second country to implement this system to indicate the main ingredients like fat, salt and sugar content in the food items [3]. France made use of an evidence based system to study different labelling systems to arrive at the best one [3]. By implementing this system, the World Health Organization will keep a check on the growing number of diet related diseases in the Europe region [3]. Europe being the largest consumer of cheese wants to regulate the ingredients that go into the manufacturing process so that people are well informed about their food choices [3].

5.1 Nutrition Grade Prediction as a Big Data Problem

We build a predictive classification model to predict the food nutrition grade based on the ingredients of the food. The goal is to apply various machine learning algorithms to the problem at hand, measure the prediction accuracy to compare and contrast the different algorithms, and arrive at the best algorithm that suits the given data and the problem. This problem can be solved using Big Data and Machine Learning techniques given the size and the complexity of the data.

6 MACHINE LEARNING

Machine Learning is a field in which we train computers in a way that they can learn from the input data [5]. The ideology is that computers use the training data that is made available to them, learn from it, build a model and use this experience to build knowledge that can be applied on new unseen data [5]. A wonderful example to demonstrate machine learning is the application to detect spam emails where the machine builds knowledge from previously seen emails which are marked as spam, checks new emails to see if they match the historic spam emails and label them as spam or non-spam [5].

6.1 Types of Machine Learning Algorithms

There are primarily two types of machine learning algorithms which are descriptive models and predictive models [5]. A *Descriptive Model* is described as the analysis done and insights gained from slicing and dicing the data in new and interesting ways [5]. One example of descriptive model is patter discovery that is often used in market basket analysis where transnational purchase details are analyzed [5]. A *Predictive Model* on the other hand involves predicting one value using one or more variables [5]. The learning algorithms tried to build a model that captures the relationship between a response variables and dependent target variables [7].

6.2 Types of Learning

Unsupervised Learning is the process where there is no explicit training data to learn from, so there is simply no mechanism where the machine can learn from previously available data [5]. The same email example can be looked at in a different way where we now want to do anomaly detection in emails [5]. Here the main goal is to detect unusual messages from the bunch of messages and we do not have experience of previous data [5].

Supervised Learning in contrast is the process of gaining knowledge or expertise from the training data which can be applied to future unseen data [5]. Here the model is first trained by using a bulk of training examples and this model is applied against testing data to measure the accuracy [5]. The variable that we need to predict is identified which is called the response variable and the variables that are used to predict the response variables, called the predictor variables are identified [5]. If the existing variables are not sufficiently giving the accuracy that is expected, a method called feature engineering is done where new variables are derived by combining existing variables [5].

7 PREDICTION ANALYSIS

Prediction analysis is the process of working on a large data set using a combination of statistical, data mining and machine learning algorithms to predict the outcome based on past data [5]. There are primarily two types of prediction analysis in machine learning, namely regression and classification [7]. In regression, we try to predict a continuous variable from the predictor variables [7]. A good example of regression is to predict the housing prices from different parameters like year of construction, location, amenities, number of bed rooms etc [7]. Here the response variable is continuous and it is not predefined [7]. Classification on the other hand tries to predict a categorical variable in which we assign each record with a predefined label or a class [7].

Classification is the task of assigning each data record to a predefined class [7]. In machine learning, classification is categorized as a supervised learning technique [7]. This problem has applications in various fields like spam detection, medical applications, astronomy and banking to identify fraudulent transactions from genuine transactions [7]. It is the task as coming up with a model which is essentially a function that maps every data record to a class label [7].

The task at hand is a classification problem since we are trying to predict the food nutrition grade of the products based on the ingredients that go into the product. For this problem we are considering only the data for the country France, since the nutrition grade is available for most food products from the country. Another reason is that France is the first countries in the region to come up with the idea of adding a color coded label to the food products mentioning the nutrition grade. In the subsequent sections we discuss the machine learning techniques used to solve this problem.

7.1 K Nearest Neighbors

7.1.1 Overview. Some of the classification algorithms in machine learning work on the principle of eager learning that involves a two step process where first a model is built from the training data and the model is applied on testing data [7]. In contrast, K nearest neighbors is a lazy learning algorithm where the process of modeling the training data is not done until the test examples are classified [7]. *Rote Classifier* is a good example of lazy learning algorithm memorizes the entire training data to perform classification but has the drawback of not being able to map every test example against the training example [7]. K nearest neighbors algorithm overcomes this drawback by finding all the records that are closest or nearest to the training records [7].

The nearest neighbour puts each attribute list as a data point in the n-dimensional space, given n the number of attributes [7]. Once we have the training examples, we take each test example and compute its distance to the training example classes and assign a class label [7]. Any of the popular distance measures among Euclidean distance, Manhattan distance, Minkowski distance and Mahalanobis distance can be used [7]. The k denotes the k closest points to the test example [7]. Figure 6 shows the structure of the data [7].

[Figure 6 about here.]

7.1.2 Support in Python. KNeighborsClassifier is available in the scikit learn python library.

7.2 Logistic Regression

Logistic regression or logit regression is a special type of regression analysis where the response variable that we need to predict is a categorical variable [7]. Typically logistic regression models the response variable to take two values 1 or 0, pass or fail, win or lose [7]. Logistic regression that takes more than two values for the response variable is called multinomial logistic regression [7]. Here the probability of the response variable to take a categorical value is modelled as a function of the predictor variables [7].

Like a lot of machine learning algorithms, logistic regression works by making a lot of assumptions which should be taken care as a part of the data cleaning and transformation process [5]. It does not assume a linear relationship between the response variables and predictor variables [5]. Since it applies a log transformation on the predicted probabilities, it can handle a variety of relationship between the predictor variables [5]. If the predictor variables are multivariate normal, the algorithm achieves best result although it works even if they are not [5]. Stepwise method must be used in the logistic regression to ensure that we are neither overfitting nor underfitting the data [5]. A very important assumption to be noted in logistic regression is that the each attribute list must be independent, in the sense the data records must not be derived from a before-after setup experiment [5]. It also requires a decently large sample size to work on [5].

7.2.1 Support for Python. LogisticRegression is available in the scikit learn python library.

7.3 Random Forest Classifier

Random forest is a ensemble classification algorithm which is very powerful [7]. Ensemble method is a special process to improve the accuracy of the prediction [7]. The classification algorithms we have seen so far predict the response variable using a single classifier on the test data but ensemble methods use multiple classfiers in tandem and aggregate the predictions to boost the accuracy by a huge margin [7]. Using a combination method, the ensemble method derives a set of base classifiers from the training data and on each iteration takes a vote of all the base classifiers to arrive at a result [7].

Random forest is an ensemble method which works very well for classification problems [7]. It combines the predictions made by multiple classifiers where each classifier independently works on the training data and casts its vote [7]. Unlike methods like AdaBoost which generates values based on independent random vectors using a varied probability distribution, random forest generates values based on fixed probability distribution [7].

7.3.1 Rationale for Random Forest. Consider an example, where we have 25 base classifiers and each base classifier has an error rate of 0.35 [7]. As discussed, the random forest takes the majority vote given by the base classifiers [7]. The model makes a wrong prediction if half or more base classifiers predict wrong, if not the accuracy is improved with an error rate of 0.06 which is far better than using just a single classifier [7].

7.3.2 Support for Python. RandomForestClassifier is available in the scikit learn python library.

8 EXPERIMENTS AND RESULTS

In this section, we will introduce the algorithm along with the details of experiments and methodology for predicting the nutrition grade of food products in France.

8.1 Algorithm

The problem at hand is to correctly identify the nutrition grade of the food item. The possible labels are, *a* to *e*, with *a* being the best and *e* being the worst grade for a food item. For this task, we have used machine learning techniques that help in predicting the label of each food item. Before getting into the details of each step of the method, we first present a concise version of the algorithm used for this task:

(1) Select all the records for country, France. Drop records where nutrition grade is not populated.
(2) Separate the predictors from the response variable in order to perform data cleaning and data transformation steps.
(3) Check for missing values in the predictors obtained in the step above. Drop columns with more than 60% missing values.
(4) Impute the missing values with 0 for remaining columns.
(5) After imputing the missing values, standardize all the numerical predictors using standard scaler.
(6) Check for the correlation between different numerical predictors. Drop one predictors from each pair of predictors that show high correlation.
(7) Combine the pre-processed predictors and the response variable in a single dataframe.
(8) Divide the data obtained in step above into training and test data using stratified sampling.
(9) Train different classifiers on the training data and check the performance of each classifier on the test data.

8.2 Data set

For the classification problem, we selected the records for country France.

Number of examples: 123,961
Number of variables: 12
Response variables: *Nutrition Grade*
Predictor variables: *Energy per 100g, Fat per 100g, Saturated Fat per 100g, Carbohydrates per 100g, Sugars per 100g, Fiber per 100g, Proteins per 100g, Salt per 100g, Trans-fat per 100g, Sodium per 100g*

8.3 Python Packages Used

The following Python packages were used to solve the classification problem:

- Pandas: Provides high performance data structures for data analysis and data munging
- Matplotlib: Plotting library that helps embedding plots into applications using GUI
- Seaborn: Visualization package based on matplotlib used for drawing high level statistical graphics
- Scikit-learn: Toolbox with solid implementation of machine learning and other algorithms
- Scipy: Package that supports scientific computing with modules for linear algebra and integration

8.4 Data Cleaning

8.4.1 Step 1: Data Sparsity. Data sparsity refers to the situation where a lot of attributes have missing values which is an advantage in some cases because you only need to store and analyze the data that is available to you and save on computation time and storage [7]. We first check the data value counts for each country. United States, France, Switzerland, Germany and Spain come as the top 5 countries with most data. Since the food nutrition grade was implemented in France, it has most products with the data available so for this classification problem we use the data filtered on France.

8.4.2 Step 2: Handling Missing Values. Missing values is a common scenario and they can be handled in different ways. You could choose to eliminate the data objects with missing values but at the expense of missing some critical analysis [7]. Estimating the missing values is also a good way to handle them, especially when the data comes from time series etc, where you could possibly interpolate the missing values from the ones that are closer to it [7]. Ignoring the missing values is another technique which can be applied for tasks like clustering where the similarity can be calculated using the attributes other than the missing ones [7].

The data set was first analyzed to check the missing values in all the columns. The threshold limit has been set at 60 percent. All the columns with missing values more than 60 percent were removed from the analysis to make the result more consistent. Once the columns were removed, the data set has to be re-indexed to maintain the order. Only the columns that are important for the prediction task has been retained from the original data set. In this case, all the ingredients which are primarily the predictor variables were included. The missing values in the response variable also need to be taken care of. Removing the records with missing values for the response variable proved to be the best option after trying out various things.

Imputation was used to handle the null values in the predictor variables. Imputation can be done in a variety of ways with by imputing the missing values by calculating the mean and the mode or just replacing them with 0. Since all the predictor variables have numeric values, all the null values have been replaced with 0. To ensure the imputation process has been done correctly, the sum of missing values is calculated.

8.4.3 Step 3: Outlier Treatment. Outliers are data objects with quite distinct characteristics from the other data records [7]. There is a considerable difference between anomalies and outliers, where anomalies refer to data records that have bad data which is noise and need to be ignored, anomalies often contain interesting aspects and can lead to some good analysis [7]. In applications like *Fraud Detection*, anomalies could be of utmost importance [7]. The outliers in the data have been looked at by using box plots and have been handled as a part of the data cleaning process.

8.5 Exploratory Data Analysis

For exploratory data analysis, we used the Seaborn package along with Matplotlib for visualizations. The measure of spread that is the range and variance of the values is a good way to understand the different aspects of the predictor variables. Box-plots are a method of visualization to look at the distribution of values for a numerical attribute [7]. The box plots show the percentiles where the lower and upper ends of the box indicate 25th and 75th percentile, the line inside the box indicates the 50th percentile, the tails indicate the 10th and 90th percentile respectively [7].

8.5.1 Bi-variate box-plots. Bi-variate box-plots go beyond univariate box plots by showing the relationship between the predictor variable and the response variable [7]. We look at the bi-variate box-plots for each of the important predictor variables namely saturated fat, polyunsaturated fat, sugars and salt and the response variable which is the nutrition grade. Figure 7 shows the bi-variate box plots.

[Figure 7 about here.]

By looking at the box plots, we can understand some important aspects of how the response variable is related to the predictor variables. We see that as the saturated fat content increases, the food grade decreases and as the polyunsaturated fat content increases the nutrition grade is better. When the sugar levels increase the health quotient of the food comes down and the energy levels behave in a interesting manner where the energy for the nutrition grade A is higher, the energy slightly increases with the nutrition grade. While increase in energy does not necessarily imply that the nutrition quality is high because there are a lot of instant energy foods that have a lot of additives but they are often rated low when it comes to health.

8.5.2 Correlation. Correlation between data objects is the measure of the linear relationship between the attributes of the object that are continuous variables [7]. Correlation analysis is the process of finding of the correlations between the different predictor variables and helps find collinearity problem [5]. The relationship could be either linear or non-linear given the data [7]. The correlation coefficient can range anywhere between -1 and 1, where 1 indicates a positive correlation and -1 indicates negative correlation [5]. Correlation plot visually shows the correlation coefficient between the variables in a nicely laid out plot. Figure 8 shows the correlation plot.

[Figure 8 about here.]

By looking at the correlation plot, we can see that sugars, fat, energy are positively in correlation with the nutrition grade. They will play an important role in the prediction algorithm. Also, sodium and salt are highly correlated with each other and this may lead to collinearity problem if not handled. Collinearity is the state where the independent variables are highly correlated with each other which can add a lot of noise to the data [6]. Some of the problems because of collinearity are that the regression coefficient may not be estimated correctly and also makes it very difficult to explain the response variables using the predictor variables [6]. So we remove sodium from the predictor variables and proceed to the next step.

8.5.3 Data Transformation. Data transformation refers to the transformation that is applied to the variables [7]. For each data object, we apply a transformation function to all the attributes of the object to ensure that the attributes do not have a lot of variance in the data [7]. This process is also called standardization since we are applying a standard function to make sure all the attributes fall within a given range [7]. There are different methods that can be applied to achieve scaling namely log transformation, absolute value, square root transformation [7].

We use the method called normalization where all the values fall in between the range 0 and 1. To achieve this, we use the preposessing package from sklearn which provides utility functions and transformer classes to change raw data into a standard representation. A lot of machine learning algorithms work well on standard data. If some of the variables have extreme values, they might dominate the model function and might disturb the estimation parameter.

There was a massive improvement in the prediction accuracy of the algorithms before and after data scaling which proves the importance of data standardization with respect to machine learning algorithms.

8.6 Data Sampling

In a supervised machine learning approach, the model is trained on one sample of the data and later tested on a different sample of the data. Thus, in order to test the performance of the nutrition grade classifier, the data for the country France was divided into two samples, training and testing. There are various ways to achieve this split or sampling of the data. Some of these sampling methods are:

- Simple Random Sampling: This is one of the simplest sampling techniques. In this technique, every data point has an equal chance of being selected. In other words, it works similar to a lottery system where every outcome has an equal probability. The biggest advantage of this technique

is the ease of implementation and its unbiased nature while generating the sample. However, random sampling might not always result in a sample that can represent the true population. It generally works well when we have huge data to sample from.

- Stratified Sampling: This technique is a more sophisticated method of sampling data. Stratified sampling generates a sample such that the proportion of each class in the sample is same as that in the true population. In this technique, the entire population is divided into groups or strata. The next step is to randomly select data points from each strata such that the final sample has the same proportion for each strata as that present in the true population. Thus, the sample generated by this technique is a good representative of the true population. Stratified sampling is a very useful technique when the classes in the data are highly imbalanced.

For our classifer, we chose to divide the data for France into training and test samples using stratified sampling technique. The strata or groups were created based on the response variable, i.e., food grade. This ensured that the training and test data had the same proportion for each food grade.

8.7 Data Modeling

Once the data was divided into training and test data, the next step was to train different classifiers and tune their respective parameters for better accuracy. We implemented three different models for classifying the food grade. Each of these models along with their parameters are:

- K Nearest Neighbors (kNN): For kNN, the grade of a food item in test data is classified by first finding the k most similar food items in the training data. It then takes vote (food grade label) from each of these neighbors and based on the majority vote, the food item in the test data is assigned a food grade. Thus, one of the most important parameter for kNN is k, i.e., the number of neighbors to consider from the training data. We tried different k values and found that k=3 gave the best accuracy.
- Logistic Regression: For logistic regression, one of the important paramaters is penalty. This parameter specifies the kind of regularization to be applied. This parameter can take two possible values, l_1 regularization and l_2 regularization. Both these values penalize high magnitude of the co-efficients of the predictors in order to prevent the model from over-fitting. For our model, we have used l_2 regularization as it works well even in the presence of highly correlated features.
- Random Forest: For random forest, there are many parameters, such as the number of trees in the forest, the maximum depth of the trees, maximum number of features to consider at each split, the minimum number of samples required in a sub-tree to qualify for further split, the minimum number of samples required to qualify as a leaf node, etc. For our data, we have kept most of the models at their default values except for the number of estimators or trees in the forest. We have set this value to 100 as the classifier produced very high accuracy with 100 trees in the forest.

8.8 Evalution Metrics and Results

There are various evalution metrics for assessing the performance of classifiers. Some of these evaluation metrics are [4]:

- Accuracy: This metric gives the proportion of the total number of correctly classified instances
- Precision: This gives the proportion of the true positive instances from the total instances classified as positive
- Recall: This gives the proportion of the positive instances that are correctly classified
- F-Measure: This gives the harmonic mean between precision and the recall values
- Confusion Matrix: This is a useful way of checking the accuracy of the classifier. It clearly shows the number of instances correctly classified for each label. Thus, if we know that the classes in the data are not well-balanced, it's always a good idea to check the confusion matrix along with accuracy. Consider a case where 95% of the instances belong to class A and only 5% of the instances belong to class B. If a classifier is trained on a dataset with such imabalance, there is a high chance that the classifier would return label A for each test instance. The classifier would be still be able to correctly classify 95% of the test instances resulting in 95% accuracy. Thus, this is a case where accuracy can be misleading and thus a quick look at the confusion matrix can help understand the problem with the classifer. For such a case, the confusion matrix will clearly show that all the instances of the minority class, B, have been misclassified.

For our model, we used accuracy as well as confusion matrix for evaluating the results. The confusion matrix did not show any serious issues for any of the classifiers. The accuracy for each of the three classifiers was:

(1) Logistic Regression: With l_2 penalty, the accuracy of logistic regression was 78.9%. Figure 9 shows the confusion matrix.

[Figure 9 about here.]

(2) K Nearest Neighbors: With k as 3, the accuracy of kNN was 95.74%. Figure 10 shows the confusion matrix.

[Figure 10 about here.]

(3) Random Forest: With number of trees as 100, the accuracy of random forest classifier was 99.68%. Figure 11 shows the confusion matrix.

[Figure 11 about here.]

Thus, we obtained the best results with Random Forest classifier.

9 CONCLUSION

Analysis of food content is very important in today's world as most of the companies try to fool consumers by labelling their product as low-fat. It's important for the consumers to know the true

nutrition grade while purchasing any food item. Thus, we tried to analyze the nutrition grade based on the composition of nutritions in the food items. We tried to build a model that labels a food item purely on the basis of its nutrients without any bias such as the production company or the brand name. For accurate labeling, we applied different data cleaning and data tranformation techniques. With this transformed data, we tried various machine learning models. We got the best results using random forest classifier which was able to accurately label 99% of the food products. Since the model is trained only for France, as part of future work, we can try and scale our model for different countries. However, to achieve similar results for other countries, we need to collect more data. The current data has many missing values for countries other than France. Once we collect enough data for these countries, we can also try and implement more sophisticated models like neural networks in future.

ACKNOWLEDGMENTS

This project was undertaken as a part of the course objective for I523: Big Data Applications and Analytics at Indiana University, Bloomington. We would like to thank Prof. Gregor von Laszewski and all the TAs for their help, support and suggestions.

REFERENCES

[1] American Heart Association. 2017. Dietary Fats. Webpage. (March 2017). https://healthyforgood.heart.org/eat-smart/articles/dietary-fats

[2] Alejandro Cifuentes. 2012. Food analysis: present, future, and foodomics. *ISRN Analytical Chemistry* 2012 (2012), 16.

[3] World Health Organization Europe. 2017. Labelling systems to guide consumers to healthier options. Webpage. (March 2017). http://www.euro.who.int/en/countries/france/news/news/2017/03/france-becomes-one-of-the-first-countries-in-region-to-recommend-colour-coded-front-of-pack-nutrition-labelling-system

[4] M Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.

[5] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, New York, USA.

[6] Statistics Solutions. 2017. Multicollinearity. Webpage. (March 2017). http://www.statisticssolutions.com/multicollinearity/

[7] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining.* Pearson, Boston, USA.

[8] Karthik Vegi and Nisha Chandwani. 2017. Code base - Analysis on food products around the world. github. (Dec. 2017). https://github.com/bigdata-i523/hid231/tree/master/project/code

List of Figures

1 Top 5 countries [8] 9
2 Top 5 countries with most fat content [8] 10
3 Top 5 countries with most saturated fat content [8] 11
4 Top 5 countries with most trans-fat content [8] 12
5 Top 5 countries with most cholesterol content [8] 13
6 K nearest neighbors algorithm[7] 14
7 Bi-variate box plots [8] 14
8 Correlation Plot [8] 15
9 Confusion matrix for Logistic Regression [8] 16
10 Confusion matrix for K Nearest Neighbors [8] 17
11 Confusion matrix for Random Forest [8] 18

Figure 1: Top 5 countries [8]

9

Figure 2: Top 5 countries with most fat content [8]

Figure 3: Top 5 countries with most saturated fat content [8]

Figure 4: Top 5 countries with most trans-fat content [8]

Figure 5: Top 5 countries with most cholesterol content [8]

Algorithm 5.2 The k-nearest neighbor classification algorithm.

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test example $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

Figure 6: K nearest neighbors algorithm[7]

Figure 7: Bi-variate box plots [8]

Figure 8: Correlation Plot [8]

Figure 9: Confusion matrix for Logistic Regression [8]

Figure 10: Confusion matrix for K Nearest Neighbors [8]

17

Figure 11: Confusion matrix for Random Forest [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
==========================================

        [2017-12-04 22.00.55] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
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 2.0s.
../README.yml
  35:16     error    trailing spaces  (trailing-spaces)
  36:81     error    line too long (599 > 80 characters)  (line-length)


================================================================================
Compliance Report
================================================================================


name: Vegi, Karthik
hid:  231
paper1: Oct 29 17 100%
paper2: 100%
project: 99%
```

```
yamlcheck
--------------------------------------------------------------------------------


wordcount
--------------------------------------------------------------------------------

(null)
wc 231 project (null) 5947 report.tex
wc 231 project (null) 5938 report.pdf
wc 231 project (null) 222 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
--------------------------------------------------------------------------------

69: We then display the top 5 countries as a pie-chart and the 5
    countries are namely United States, France, Switzerland, Germany,
    and Spain as shown in Figure \ref{fig:Fig7}.
71: \begin{figure}
72: \includegraphics[width=1.0\textwidth]{images/fig7.png}
74: \label{fig:Fig7}
80: The top 5 countries with most fat content in the food items are
    Serbia, United States, Switzerland, Germany, and Sweden as shown
    in Figure \ref{fig:Fig8}.
```

```
82: \begin{figure}
83: \includegraphics[width=1.0\textwidth]{images/fig8.png}
85: \label{fig:Fig8}
88: The top 5 countries with most saturated fat content in the food
    items are Serbia, United States, Germany, France and Switzerland
    as shown in Figure \ref{fig:Fig9}
90: \begin{figure}
91: \includegraphics[width=1.0\textwidth]{images/fig9.png}
93: \label{fig:Fig9}
97: The top 5 countries with most trans-fat content in the food items
    are United States, Brazil, Canada, Australia, Russia, and Serbia
    as shown in Figure \ref{fig:Fig10} \\
99: \begin{figure}
100: \includegraphics[width=1.0\textwidth]{images/fig10.png}
102: \label{fig:Fig10}
105: The top 5 countries with most cholesterol content in the food
     items are United States, Canada, Portugal, Brazil, France, and
     Italy as shown in Figure \ref{fig:Fig11} \\
107: \begin{figure}
108: \includegraphics[width=1.0\textwidth]{images/fig11.png}
110: \label{fig:Fig11}
146: The nearest neighbour puts each attribute list as a data point in
     the n-dimensional space, given n the number of attributes
     \cite{book-tan}. Once we have the training examples, we take each
     test example and compute its distance to the training example
     classes and assign a class label \cite{book-tan}. Any of the
     popular distance measures among Euclidean distance, Manhattan
     distance, Minkowski distance and Mahalanobis distance can be used
     \cite{book-tan}. The k denotes the k closest points to the test
     example \cite{book-tan}. Figure \ref{fig:Fig1} shows the
     structure of the data \cite{book-tan}.
148: \begin{figure}
149: \includegraphics[width=1.0\textwidth]{images/fig1.png}
151: \label{fig:Fig1}
226: \subsubsection{Bi-variate box-plots} Bi-variate box-plots go
     beyond uni-variate box plots by showing the relationship between
     the predictor variable and the response variable \cite{book-tan}.
     We look at the bi-variate box-plots for each of the important
     predictor variables namely saturated fat, polyunsaturated fat,
     sugars and salt and the response variable which is the nutrition
     grade. Figure \ref{fig:Fig2} shows the bi-variate box plots. \\
228: \begin{figure}
229: \includegraphics[width=1.0\textwidth]{images/fig2.png}
231: \label{fig:Fig2}
237: Correlation between data objects is the measure of the linear
     relationship between the attributes of the object that are
```

```
     continuous variables \cite{book-tan}. Correlation analysis is the
     process of finding of the correlations between the different
     predictor variables and helps find collinearity problem
     \cite{book-shai}. The relationship could be either linear or non-
     linear given the data \cite{book-tan}. The correlation
     coefficient can range anywhere between -1 and 1, where 1
     indicates a positive correlation and -1 indicates negative
     correlation \cite{book-shai}. Correlation plot visually shows the
     correlation coefficient between the variables in a nicely laid
     out plot. Figure \ref{fig:Fig3} shows the correlation plot. \\
239: \begin{figure}
240: \includegraphics[width=1.0\textwidth]{images/fig3.png}
242: \label{fig:Fig3}
287: \item Logistic Regression: With $l_2$ penalty, the accuracy of
     logistic regression was 78.9\%. Figure \ref{fig:Fig4} shows the
     confusion matrix. \\
289: \begin{figure}
290: \includegraphics[width=1.0\textwidth]{images/fig4.png}
292: \label{fig:Fig4}
295: \item K Nearest Neighbors: With k as 3, the accuracy of kNN was
     95.74\%. Figure \ref{fig:Fig5} shows the confusion matrix. \\
297: \begin{figure}
298: \includegraphics[width=1.0\textwidth]{images/fig5.png}
300: \label{fig:Fig5}
303: \item Random Forest: With number of trees as 100, the accuracy of
     random forest classifier was 99.68\%. Figure \ref{fig:Fig6} shows
     the confusion matrix. \\
305: \begin{figure}
306: \includegraphics[width=1.0\textwidth]{images/fig6.png}
308: \label{fig:Fig6}

figures 11
tables 0
includegraphics 11
labels 11
refs 11
floats 11

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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

72: \includegraphics[width=1.0\textwidth]{images/fig7.png}

83: \includegraphics[width=1.0\textwidth]{images/fig8.png}

91: \includegraphics[width=1.0\textwidth]{images/fig9.png}

100: \includegraphics[width=1.0\textwidth]{images/fig10.png}

108: \includegraphics[width=1.0\textwidth]{images/fig11.png}

149: \includegraphics[width=1.0\textwidth]{images/fig1.png}

229: \includegraphics[width=1.0\textwidth]{images/fig2.png}

240: \includegraphics[width=1.0\textwidth]{images/fig3.png}

290: \includegraphics[width=1.0\textwidth]{images/fig4.png}

298: \includegraphics[width=1.0\textwidth]{images/fig5.png}

306: \includegraphics[width=1.0\textwidth]{images/fig6.png}

passed: False

below_check
--------------------------------------------------------------------------------


bibtex
--------------------------------------------------------------------------------


label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
```

```
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib

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
```

Big DATA IN RAIN WATER HARVESTING

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

Rainwater harvesting is the accumulation and deposition of rainwater for reuse on-site, rather than allowing it to run off. Rainwater can be collected from rivers or roofs, and in many places, the water collected is redirected to a deep pit (well, shaft, or borehole), a reservoir with percolation, or collected from dew or fog with nets or other tools. Its uses include water for gardens, livestock, irrigation, domestic use with proper treatment, indoor heating for houses, etc. The harvested water can also be used as drinking water, longer-term storage, and for other purposes such as groundwater recharge.Rainwater harvesting is one of the simplest and oldest methods of self-supply of water for households usually financed by the user. [4]. Rainwater harvesting is also used to tackle the problem of water scarcity. Water scarcity caused due to pollution, global warming and overuse has become a huge threat to the existence of man. solving this simply by filtration and redistribution of water from dams and from normal rainfall, these can be augmented with rainwater harvesting systems.

2 BIG DATA IN RAIN WATER HARVESTING

2.1 Introduction

Before the subject matter of big data in rainwater harvesting is tackled it is first necessary to understand the rainwater harvesting process before the combination with big data can be explained. For the purpose of this study, the method rooftop rainwater harvesting is used. In brief, the rainwater harvesting process can be grossly oversimplified as follows:

- Analyze feasibility of installation
- Installation
- First wash
- route rainwater to storage tank
- redirect water in case of overflow

the figure 1 provides a good explanation on rainwater harvesting process

[Figure 1 about here.]

and a good article to explain the rainwater harvesting process can be found here [4]. The first wash step, in particular, is very important because it removes dust debris etc from the rooftop or else we risk contamination of water. Quite naturally we cannot allow the collected rainwater to overfill tanks in case of this we need to redirect the water to some other outlet.Big data can play a very influential role right from the feasibility analysis to re direction of water.

2.2 Big data and feasibility of installation

Big data can play a huge role in the feasibility estimation. This can be useful for both households and governments, in many countries in some states it is mandatory that each house should have a rainwater harvesting unit. In some cases, these are funded by the government and in some cases it in on the owner to do so. In the case of governments to do an analysis how can they do so, it is a huge task to go to each and every house and track roof dimensions. One easy way of going about this would be to use satellite image data. These images can then be searched for roof features and dimensions accordingly extrapolated and in such a manner the dimensions of many roofs can be obtained and a cost estimate can be obtained. To obtain the data we can use the many datasets provided by NASA or we can even use a highly zoomed street view from google maps. To extract the features we can use one of the many open CV libraries or use apply complex ML deep learning algorithms . To store this data a simple Hadoop map-reduce can be used. A more detailed study can be viewed in [2].

2.3 Big data and first wash

the importance of first wash was previously stressed upon in the introduction. it is required to wash away dust, debris, dead insects and other such contaminants. The first watch is a manual task it is dependent on the owner to redirect the first wash water elsewhere. Often most people have mistaken it to be the first wash to be the first rain, that is people waste a whole day of rain at times as first wash, or some mistakenly use small drizzles as the first and do not use the first watch properly. Big data can help in the automation of the first wash process combined with IoT. The first step would be to obtain the weather data, this can be achieved either by using the highly consolidated data obtained from the respective government's meteorology departments or the huge datasets provided by the NASA satellite. Once again Hadoop map-reduce allows for reducing

a humongous data set into more compact usable structures. From this we can perform a weather data analysis to determine if the rain will be heavy/light and its duration. From this data, we can easily determine when to perform the first watch. Assuming every rainwater harvesting unit has an IoT feature that controls the valves or water redirection one central control center can send signals to a wide area on when to perform the first watch and for how long. This should greatly reduce the amount of rainwater wasted.

2.4 Big data and water tanks

Big data and IoT can once again help in the rainwater harvesting process, there are many times water left in the tanks are not used and the water becomes stagnant with the use of devices the quality of water can be checked and the tanks can be drained. Also, it is very difficult to combine the rainwater tanks with the main water channels of the buildings since the water in the tanks is very very limited. With the help of IoT redistributing this water becomes very easy using data from other parts of the housing analysis can be made water can be redistributed accordingly. Then there is also the matter of making sure the tanks don't overflow which could lead to bursting. Using IoT the water can be tracked in a smart manner and decisions like when to reroute can be done in a smart manner. There is one more important use for big data in tanks, leakages and rusting. As previously mentioned the quality of water can be checked by its ph level using smart devices the next issue comes down to leakages. more often than not most installations are buried under the ground this is done in order to reduce the effects of weather and also so that the installation doesn't take up space. While this leads to a new set of problems the major one is that often leaks cant be detected until its too late. IoT devices can be used to alert a user that water levels are falling down way too rapidly and the user can contact servicemen in time before the next rains.

2.5 Big data and water re routing

Lastly but perhaps most important is the issue of rerouting water once the tanks get full. In the majority of the cases, the rainwater is directly diverted to the water table. While it is normally a good idea to replenish the water table in such a manner due to over exploitation from bore wells. Aside from restoring the water table, it is slowly becoming essential to recharge even the lakes and other sources of freshwater. This is becoming important because with global warming and rainfall becoming more erratic [some places receiving more rainfall than the others and others receiving way lesser] as a result we need to divert some of the harvested rainwater to other lakes/reservoirs. As to how this can be achieved we can use Big data to monitor the water levels and then decide accordingly where to route the rainwater and by how much.

3 TECH IN RAIN WATER HARVESTING

Surprisingly there is not much to write about about very few players exist in the rainwater harvesting market who aim to offer the services of big data. Part of this can be attributed to the fact there is not much data available. For example, Indian government offers highly consolidated annual precipitation data for free while this is useful to perform past estimates it however is really not enough, more detailed minute by minute data of precipitation is required in order for the above mentioned analysis to take place. Even the NASA weather data doesn't give the whole picture. This doesn't mean the data is not there but a premium is required to obtained it. there are a few players who offer smart tank service . However the tech scene is just getting warmed up and awareness of its potential is doing rounds a good article was recently released by NASA on this [1] and a few examples are [3].

4 CONCLUSION

The scope for big data in rainwater harvesting is immense but the major initative lives with the governments, rerouting water into reservoirs is not in the best interests of private contractors but they can be hired to make it so. This can create a huge job market and it will be beneficial for all involved. This also needs to be done sooner rather than later because the rate of population growth exceeds our available sources and if we want the future generations to have any resources we need to embrace technology and start protecting our assets. Also, for private contractors to approach the government with proposals it would be very useful if more relevant data was made available to the public and thus there is a need for investment in the weather department for data.Rainwater harvesting despite being one of the oldest practices of water replenishment is surprisingly behind in terms of technology advancement when we look at the progress made with solar and wind.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ashley Morrow. 2015. Using NASA Data to Show How Raindrops Could Save Rupees. *NASA* 1, 1 (Jan. 2015), 1. https://www.nasa.gov/feature/goddard/using-nasa-data-to-show-how-raindrops-could-save-rupees

[2] Robert O. Ojwang. 2015. Rooftop Rainwater Harvesting for Mombasa: Scenario Development with Image Classification and Water Resources Simulation. *Water 2017* 1, 1 (Jan. 2015), 1. http://www.mdpi.com/2073-4441/9/5/359/htm

[3] UNEP. 2017. rainwater harvesting examples. *unep* 1, 1 (Jan. 2017), 1. http://www.unep.or.jp/ietc/publications/urban/urbanenv-2/9.asp

[4] Wikipedia. 2016. Rain water harvesting. *wikipedia* 1, 1 (Jan. 2016), 1. https://en.wikipedia.org/wiki/Rainwater_harvesting

A ISSUES

DONE:

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

A.1 Assignment Submission Issues

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

A.2 Uncaught Bibliography Errors

Missing bibliography file generated by JabRef

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

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

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

Other formatting issues

A.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

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

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

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

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

Use a space after . , :

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

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

A.5 Citation Issues and Plagiarism

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

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

A.6 Character Errors

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

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

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

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.

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

A.7 Structural Issues

Acknowledgement section missing

Incorrect README file

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

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

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

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

A.8 Details about the Figures and Tables

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

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

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

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

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

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

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

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

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

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

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

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

Do not use textwidth as a parameter for includegraphics

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

e.g.

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

re

List of Figures

1 rainwater harvesting figure 5

Figure 1: rainwater harvesting figure

```
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
===========================================

        [2017-12-04 22.01.02] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.2s.

================================================================================
Compliance Report
================================================================================


name: Rahul Velayutham
hid:  232
paper1: 2017-10-29 100%
paper2: 100%
project: in progress 40%
```

```
yamlcheck
--------------------------------------------------------------------------------


wordcount
--------------------------------------------------------------------------------

5
wc 232 project 5 1768 report.tex
wc 232 project 5 2448 report.pdf
wc 232 project 5 151 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
--------------------------------------------------------------------------------

50: the figure \ref{f:water} provides a good explanation on rainwater
    harvesting process
51: \begin{figure}[!ht]
52: \centering\includegraphics[width=\columnwidth]{images/water.pdf}
53: \caption{rainwater harvesting figure}\label{f:water}

figures 1
tables 0
```

```
includegraphics 1
labels 1
refs 1
floats 1

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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

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

bibtex_empty_fields
--------------------------------------------------------------------------------



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

```
find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



================================================================================
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 the Travel Industry and its Potential in Improving Travel Accessibility

Weixuan Wang
Indiana University Bloomington
School of Public Health
Bloomington, Indiana 47405
wangweix@indiana.edu

ABSTRACT

This is my abstract

KEYWORDS

i523, HID234

1 INTRODUCTION

Big Data Applications in the Travel Industry and its Potential in Improving Travel Accessibility. Here is the introduction [1].

2 TRANSPORTATION AND BIG DATA

3 TRAVEL AND BIG DATA

4 PROMISE OF BIG DATA IN TRAVEL ACCESSIBILITY

5 CONCLUSION

Put here an conclusion. Conlcusions and abstracts must not have any citations in the section.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

```
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 chapter and pages in editor00
(There was 1 warning)

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

report.bib:22:@Inbook{las_gergor00,

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


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


latex report
===========================================

        [2017-12-04 22.01.14] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Typesetting of "report.tex" completed in 0.9s.
../README.yml
  9:81      error    line too long (86 > 80 characters)  (line-length)
  24:81     error    line too long (119 > 80 characters)  (line-length)
  25:81     error    line too long (125 > 80 characters)  (line-length)
  25:125    error    trailing spaces  (trailing-spaces)
  26:81     error    line too long (118 > 80 characters)  (line-length)
  26:118    error    trailing spaces  (trailing-spaces)
  32:28     error    trailing spaces  (trailing-spaces)
  33:81     error    line too long (91 > 80 characters)  (line-length)
  38:81     error    line too long (82 > 80 characters)  (line-length)
  38:82     error    trailing spaces  (trailing-spaces)
  39:81     error    line too long (82 > 80 characters)  (line-length)
  39:82     error    trailing spaces  (trailing-spaces)
  40:81     error    line too long (87 > 80 characters)  (line-length)
  41:81     error    line too long (86 > 80 characters)  (line-length)
  41:86     error    trailing spaces  (trailing-spaces)
```

```
  42:81     error    line too long (84 > 80 characters)  (line-length)
  42:84     error    trailing spaces  (trailing-spaces)
  43:81     error    line too long (86 > 80 characters)  (line-length)
  43:86     error    trailing spaces  (trailing-spaces)
  45:1      error    trailing spaces  (trailing-spaces)
  51:12     error    too many spaces after colon  (colons)
  51:81     error    line too long (108 > 80 characters)  (line-length)


================================================================================
Compliance Report
================================================================================



name: Weixuan Wang
hid:  234
paper1: Oct 22 2017 100%
paper2: Nov 9 2017 100%
project: 70%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

1
wc 234 Project 1 125 content.tex
wc 234 Project 1 131 report.pdf
wc 234 Project 1 144 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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors
```

```
22: las_gergor00: do not use underscore in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty chapter and pages in editor00
(There was 1 warning)

bibtex_empty_fields
--------------------------------------------------------------------------------


entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

3: author =         "",

15: chapter =        "",

16: pages =          "",

17: number =         "",

18: type =   "",

19: month =          "",

20: note =   "",

36: chapter =        "",

37: pages =          "",

38: number =         "",

39: type =   "",

40: month =          "",
```

```
41: note =   "",

passed: False

ascii
--------------------------------------------------------------------------------



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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True


bibtext report
===========================================

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "editor00"
(There was 1 warning)

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


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


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

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

        [2017-12-04 20.20.17] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
p.1   L62   : [editor00] undefined
p.1   L89   : 't:mytable' undefined
Empty `thebibliography' environment.
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
There were undefined citations.
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
There were undefined references.
Typesetting of "report.tex" completed in 1.3s.

==================================================================================
Compliance Report
==================================================================================


name: Ahmed, Tousif
hid:  237
paper1: 100%, October 27, 2017
paper2: 100%, Nov 6, 2017
project: 10%, Dec 4, 2017

yamlcheck
----------------------------------------------------------------------------------


wordcount
----------------------------------------------------------------------------------

(null)
wc 237 project (null) 518 report.tex
wc 237 project (null) 1163 report.pdf
wc 237 project (null) 50 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
--------------------------------------------------------------------------------

67: In Figure \ref{f:fly} we show a fly. Please note that because we
    use
73: \begin{figure}[!ht]
74: \centering\includegraphics[width=\columnwidth]{images/fly.pdf}
75: \caption{Example caption}\label{f:fly}
90: or generate them by hand while using the provided template in
    Table\ref{t:mytable}. Not ethat
93: \begin{table}[htb]
96: \label{t:mytabble}

figures 1
tables 1
includegraphics 1
labels 2
refs 2
floats 2

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

```
Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "editor00"
(There was 1 warning)

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------

15: note            = "",
```

```
passed: False

ascii
--------------------------------------------------------------------------------



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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

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, Morningstar, or others 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 [5]. 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 [2].

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 did stuff... NASDAQ began as an all-electric equities exchange in 1971 and today provides trading, technology, and information services for financial markets. Today [8].

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 [4].

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 [1].

An Application Programming Interface (API) acts as the middle-man between the requesting service and the preforming 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.

1.1 Pharmaceutical Sector

The pharmaceutical industry

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 to create a system that consistently predicts or aids in identifying drug targets.

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 [9]. 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 [3].

Clinical trials have four defined phases. Phase I trials identify how well a drug is tolerated by determining the maximally tolerated doe (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 [3].

Clinical trial designs have been very slow to evolve due to

1.3 Established Analyst Ratings

1.4 Data Resources

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 http://www.nasdaq.com/quotes/. NASDAQ provides API services for subscribers starting at $5,000 per year [7]. 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 [10]. The package can be installed through Git by cloning the Git directory where the package is available: https://github.com/lukaszbanasiak/yahoo-finance.git. To install the python package without Git the tape archive can be downloaded from https://pypi.python.org/pypi/yahoo-finance. While Yahoo! Finance is a great resource, the API does not function consistently. To install the python package without Git, the tape archive can be downloaded from https://pypi.python.org/pypi/yahoo-finance. Tape archives allow for compression of multiple files which can be restored to their original format using the tar command in the command line [6]. 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
```

2 METHODS

2.1 Data Collection

Data collection was initiated with the use to Google Alerts. Google allows for alerts to be configures from https://www.google.com/alerts. Gmail users can configure these alerts to be sent through email when news or other types of articles 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 this data. 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. Once credentials are received 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 quickstart file which facilitates the first authentication run. Running this quick start guide will open a browser window and prompt the user to log into a Gmail account. The user then accepts the authorization and can run the Gmail API from command line or other compilers.

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

The Python code that calls the Gmail API and writes a csv from the data starts, after calling all needed libraries, with defining the scope of the authorization. Google mail can be opened with a Read-only 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 labe lID "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 atleast 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 fron "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 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.

[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.

2.2 Data Analysis

3 RESULTS

3.1 Comparison to Established Analyst Ratings

4 CONCLUSION

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 to write this paper.

REFERENCES

[1] 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

[2] F. Braudel. 1982. *Civilization and Capitalism, 15th-18th Century, Vol. II: The Wheels of Commerce.* University of California Press. https://books.google.com/books?id=WPDbSXQsvGIC

[3] L.M. Friedman, C. Furberg, and D.L. DeMets. 1998. *Fundamentals of Clinical Trials.* Springer. https://books.google.com/books?id=yzxT0Zh3X3IC

[4] Thomas Hellstrom and Kenneth Holmstrom. 1997. *Predict the stock market.* techreport. Department of Mathematics and Physics, Malardalen University, Sweden.

[5] Investopedia. 2017. Stock Market. Website. (09 2017). https://www.investopedia.com/terms/s/stockmarket.asp?lgl=rira-layout

[6] LINFO. 2006. The tar Command. website. (07 2006). http://www.linfo.org/tar.html

[7] NASDAQ. 2017. NASDAQ DataOnDemand Subscription Plans. Website. (2017). https://www.nasdaqdod.com/Shop/ProductConfig.aspx?product=webservices&service=NASDAQDataOnDemand

[8] NASDAQ. 2017. NASDAQ's Story. website. (2017). http://business.nasdaq.com/discover/nasdaq-story/index.html

[9] S.J. Pocock. 2013. *Clinical Trials: A Practical Approach.* Wiley. https://books.google.com/books?id=TxbTBQAAQBAJ

[10] Python.org. 2016. yahoo-finance 1.4.0. Website. (11 2016). https://pypi.python.org/pypi/yahoo-finance

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. 5

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 atleast 10%. It also prints the stocks that increased by atleast 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. 5

```
your code here
```

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

```
your code here
```

Figure 2: This Python script takes in the Date, Stock Ticker Symbol, and Snippet from the Google API .csv that was produced using both manual mining of the stock symbols and the python script provided for getting the Date and Snippet from Gmail. This code returns a modified .csv which lists an "L" for stocks that did not increase by 10% in five days and a "W" for stocks that increased by atleast 10%. It also prints the stocks that increased by atleast 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.

```
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 address in Braudel
Warning--empty address in Friedman
Warning--can't use both author and editor fields in Pocock
Warning--empty address in Pocock
(There were 4 warnings)

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


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


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


latex report
===========================================

        [2017-12-04 22.01.41] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.1s.
../README.yml
  69:81     error    line too long (83 > 80 characters)  (line-length)


=================================================================================
Compliance Report
=================================================================================


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

```
paper2: 100%
project: 90%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

(null)
wc 313 project (null) 2437 report.tex
wc 313 project (null) 2632 report.pdf
wc 313 project (null) 410 report.bib

find "
--------------------------------------------------------------------------------

108: Data collection was initiated with the use to Google Alerts.
     Google allows for alerts to be configures from
     https://www.google.com/alerts. Gmail users can configure these
     alerts to be sent through email when news or other types of
     articles 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 this data. On slow days,
     only a couple alerts were received.  Only very infrequently were
     no messages received.

158: $ PATH="$PATH:/c/Python27"

passed: False

find footnote
--------------------------------------------------------------------------------

passed: True

find input{format/i523}
--------------------------------------------------------------------------------

4: \input{format/i523}

passed: True
```

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

passed: False

floats
--------------------------------------------------------------------------------

130: \begin{figure}[htb]
134: \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}
137: Figure \ref{c:googleapi} shows the entire code to extract Google
     Alerts data using the Google provided Gmail API.
173: \begin{figure}[htb]
177: \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 atleast 10\%. It also prints the stocks that
     increased by atleast 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}
180: Figure \ref{c:stock} shows the code to combine the data produced
     by the Google Alert mining and available historic stock price
     data.

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

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

```
Label/ref check
passed: True

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty address in Braudel
Warning--empty address in Friedman
Warning--can't use both author and editor fields in Pocock
Warning--empty address in Pocock
(There were 4 warnings)

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
--------------------------------------------------------------------------------
```

```
passed: True

ascii
--------------------------------------------------------------------------------


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

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

find newline
--------------------------------------------------------------------------------

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

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

passed: True
```

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 or 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 braches, 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 the 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 to 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
[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
[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
[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
[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
[31] Spandas Lui. 2012. How eBay Uses Big Data to Make You Buy More. (2012). www.zdnet.com
[32] Charu Manglani. 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] Chistopher Meyer. 2007. Understanding Customer Experience. (2007).
[35] Timothy Pickett Morgan. 2014. Why Hadoop is the New Backbone of American Express. (2014). www.enterprisetech.com
[36] Harikesh S. Nair, Sanjog Misra, Wlliam J. Hornbuckle IV, Ranjan Mishra, and Anand Acharya. 2016. Big Data and Marketing Analytics in Gaming: Combining Empiral Models and Field Experimentation. (2016).
[37] Raghunath Nambiar, Ruchie Bhardwaj, Adhiraaj 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
[40] Greg Petro. 2017. Amazon's Acqisition of Whole Foods is About Two Things: Data and Product. (2017). 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 Stastics. (2017).
[44] Cliff Saran. 2014. How Big Data is Powering Success for eBay's Customer Journey. (2014). 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
[48] Jeffrey Spiess, Yves T'Joens, Raluca Dragnea, Peter Spencer, and Laurent Philppart. 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

List of Figures

1 Customer Experience Journey Touchpoints 12
2 Customer Experience Journey 12
3 Digital Marketing Customer Stages Model 13

Figure 1: Customer Experience Journey Touchpoints

Figure 2: Customer Experience Journey

Figure 3: Digital Marketing Customer Stages Model

```
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
===========================================

        [2017-12-04 22.02.05] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.1s.
../README.yml
  35:81     error    line too long (1061 > 80 characters)  (line-length)
  35:1060   error    trailing spaces  (trailing-spaces)


================================================================================
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}\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}\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.jp
    g}
73: \caption{Digital Marketing Customer Stages Model}\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 refered 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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

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

bibtex_empty_fields
--------------------------------------------------------------------------------



entries in general should not be empty in bibtex


find ""
```

```
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



================================================================================
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 healthcare 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 adminstrative 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 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 ϵ 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 ϵ 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 ϵ, 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.

$$\left\{ \begin{array}{cc} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{array} \right\}$$

This is the general framework of a confusion matrix which shows the counts of each type of prediction and the accuracy of that prediciton. 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 specifity 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 unncecessary 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 appropirate 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 the 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:

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

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 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 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 code issue occurring. 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 inital 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 variabeles. 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:

$$
\begin{aligned}
X_1 &= \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \beta_k X_k \\
X_2 &= \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \beta_k X_k \\
X_3 &= \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \ldots + \beta_k X_k \\
&\ldots \\
X_k &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{k-1} X_{k-1}
\end{aligned}
$$

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 more drastic weights, 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:

$$\left\{ \begin{array}{cc} 12713 & 2 \\ 1282 & 1 \end{array} \right\}$$

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 thatn 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 hish 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 incapable 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. Furhter, 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.

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 Multicollienarity 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 Miller. 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 Huftless. 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

[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 Yahi, Riccardo Miotto, Li Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P Sengupta, Annetine Gelijns, Alan Moskovitz, 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

4 FIGURES

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

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

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
==========================================

        [2017-12-04 22.02.14] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
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 1.9s.

================================================================================
Compliance Report
================================================================================


name: Tyler Peterson
hid:  331
paper1: Oct 22 17 100%
paper2: Nov 6 17 100%
project: Dec 04 17 99%

yamlcheck
--------------------------------------------------------------------------------


wordcount
```

```
--------------------------------------------------------------------------------

12
wc 331 project 12 6923 report.tex
wc 331 project 12 7047 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
--------------------------------------------------------------------------------

161: 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.
174: 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 thatn 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.
192: \begin{figure}[!ht]
193: \centering\includegraphics[width=\columnwidth]{images/weightsl2.p
     ng}
194: \caption{Logistic Regression Weights By C-Value, L2
     Regularization}\label{f:weightsl2}
197: \begin{figure}[!ht]
198: \centering\includegraphics[width=\columnwidth]{images/weightsl1.p
     ng}
199: \caption{Logistic Regression Weights By C-Value, L1
     Regularization}\label{f:weightsl1}
202: \begin{figure}[!ht]
203: \centering\includegraphics[width=\columnwidth]{images/roccurve001
     .png}
204: \caption{ROC Curve, L1 Regularization, C =
     0.01}\label{f:roccurve001}
207: \begin{figure}[!ht]
208: \centering\includegraphics[width=\columnwidth]{images/roccurve100
     .png}
209: \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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

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

bibtex_empty_fields
--------------------------------------------------------------------------------
```

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


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



================================================================================
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 to potential can 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 nations 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 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 State by 12 to 17 percent. This equates 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 is 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 (OEDC), the United States spends 2.5 times per person than the average of OEDC 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 OEDC 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 OEDC 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 countries size and wealth as compared to other OEDC 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 OEDC countries is 77.9 years. The incidence of obstetric trauma is 9.6 per 100000 births in the United States compared to 5.7 incidents per 100000 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 100000 in comparison to the average of 236 per 100000. Thirty eight percent of the population in the United States is obese. The average obesity rate in other countries at 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, uncoordinated, multi-payer payment structure. Private insurance companies, Medicaid, and Medicare are the primary payers. An individuals 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 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 and 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 on 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 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 ste waste: operational, and clinical [3].

Operation 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 is 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 directed 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 is 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 in 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 State 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 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 [**?**]. 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 trail 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 by 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 Clincal Library Association. 2011. Genetic Testing Can Help the United States Cut Costs and Improve Care. Web page as article. (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 Healthcare Costs. Web page as article. (Feb. 2017). https://www.aeaweb.org/research/health-care-price-transparency

[3] Effros Rachel M Palar Kartika Keeler Emmett B Bentleu, Tanya. 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. (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. (2017). https://www.datameer.com/company/datameer-blog/role-big-data-preventing-healthcare-fraud-waste-abuse/

[11] Digitalist. 2016. Can Big Data Analytics Save Billions in Healthcare Costs. Web page as Article. (Feb. 2016). http://www.digitalistmag.com/resource-optimization/2016/02/29/big-data-analytics-save-billions-in-healthcare-costs-04037289

[12] Forbes. 2015. How Big Data in changing Healthcare. Web page as Article. (April 2015). https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#427274d12873

[13] Forbes. 2016. How an Electronic Health Record can Save Time, Money and Lifes. Web page. (Dec. 2016). https://www.forbes.com/sites/robertpearl/2016/12/01/how-an-electronic-health-record-can-save-time-money-and-lives/2/#4445b8275f57

[14] Harvard Business Review. 2014. How Cities are Using Analytics to Improve Public Health. Web page as article. (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. (2013). http://healthcare.mckinsey.com/big-data-revolution-us-healthcare

[19] KQED Science. 2015. How San Diego is Using Big Data to Improve Public Health. Web page as article. (Aug. 2015). https://ww2.kqed.org/futureofyou/2015/08/19/how-san-diego-is-using-big-data-to-improve-public-health/

[20] Liason. 2017. Value Based Healthcare - The patient is the Center but Data is the Key. Web page as blog. (2017). https://www.liaison.com/blog/2017/06/22/value-based-healthcare-patient-center-data-key/

[21] Managed Healthcare Execution. 2017. Five ways to reduce healthcare administrative costs. Web page as article. (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. (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 pharmaceuticla R and D. Web page as article. (Nov. 2017). https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d

[25] Pacient. 2017. How Big Data Can Improve Health Care. Web page as article. (Nov. 2017). https://pacient.care/decks/privacy-technology/health-technology/how-big-data-can-improve-healthcare

[26] PBSO News Hour. 2012. Health Costs: How the US Compares with Other Countries. Web page as Article. (Oct. 2012). https://www.pbs.org/newshour/health/health-costs-how-the-us-compares-with-other-countries

[27] Practice Fusion. 2017. EHR Adoption Rates 20 Must see stats. Web page as Article. (March 2017). https://www.practicefusion.com/blog/ehr-adoption-rates/

[28] OEDC Publishing. 2017. *Health at a Glance 2017.* OECD, Paris. http://dx.doi.org/10.1787/health_glance-2017-en

[29] Raghupahti Viju Raghupahti, Wullianallur. 2014. Big Data Analytics in Healthcare Promise and Potential. *Spinger 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. (2017). https://www.searchtechnologies.com/blog/predictive-analytics-in-healthcare

[32] Stephen B Thacker. 2003. HIPAA Privacy Rule and Public Health. *CDC* 52 (April 2003), 1–12. https://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

[33] The Common Wealth Fund. 2017. The Affordable Care Acts Payment and Delivery System reforms: A progress report. Web page as article. (Feb. 2017). http://www.commonwealthfund.org/publications/issue-briefs/2015/may/aca-payment-and-delivery-system-reforms-at-5-years

[34] The datapine blog. 2017. Nine examples of Big Data Analytics in Healthcare that Can Save People. Web page. (May 2017). https://www.datapine.com/blog/big-data-examples-in-healthcare/

[35] The Wall Street Journal. 2017. How to Research Medical Prices. Web page as article. (Nov. 2017). http://guides.wsj.com/health/health-costs/how-to-research-health-care-prices/

[36] US Department of Health and Human Resources. 2017. EHR Basics. Web page. (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

[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
Warning--string name "april" is undefined
--line 15 of file report.bib.bib
Warning--string name "june" is undefined
--line 33 of file report.bib.bib
Warning--string name "june" is undefined
--line 108 of file report.bib.bib
Warning--string name "sept" is undefined
--line 117 of file report.bib.bib
Warning--string name "sept" is undefined
--line 126 of file report.bib.bib
Warning--string name "sept" is undefined
--line 135 of file report.bib.bib
Warning--string name "sept" is undefined
--line 186 of file report.bib.bib
Warning--string name "july" is undefined
--line 249 of file report.bib.bib
Warning--string name "sept" is undefined
--line 330 of file report.bib.bib
Warning--string name "april" is undefined
--line 348 of file report.bib.bib
Warning--I didn't find a database entry for "www-google-christion"
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
```

```
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
```

```
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 40 error messages)
make[2]: *** [bibtex] Error 2

latex report
============================================

        [2017-12-04 22.02.20] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
p.5   L139  : [www-google-christion] undefined
There were undefined citations.
Typesetting of "report.tex" completed in 1.0s.

================================================================================
Compliance Report
================================================================================



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

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------
```

```
10
wc 332 project 10 8955 report.tex
wc 332 project 10 9322 report.pdf
wc 332 project 10 1551 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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
Warning--string name "april" is undefined
--line 15 of file report.bib.bib
Warning--string name "june" is undefined
--line 33 of file report.bib.bib
Warning--string name "june" is undefined
--line 108 of file report.bib.bib
Warning--string name "sept" is undefined
--line 117 of file report.bib.bib
Warning--string name "sept" is undefined
--line 126 of file report.bib.bib
Warning--string name "sept" is undefined
--line 135 of file report.bib.bib
Warning--string name "sept" is undefined
--line 186 of file report.bib.bib
Warning--string name "july" is undefined
--line 249 of file report.bib.bib
Warning--string name "sept" is undefined
--line 330 of file report.bib.bib
```

```
Warning--string name "april" is undefined
--line 348 of file report.bib.bib
Warning--I didn't find a database entry for "www-google-christion"
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3085 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
```

```
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3131 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
Too many commas in name 1 of "Bentleu,Tanya, Effros, Rachel M, Palar, Kartika, Keeler, E
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 40 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 {
--------------------------------------------------------------------------------

138: 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
     \cite{www-google-elec}.  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
     \cite{www-google-datapine}.  Productivity is increased, overlap
     care is reduced, and coordination of care is enhanced \cite {www-
```

```
    google-christian}. In general, electronic health records (EHRs)
    improve quality of care enhance patient safety, and contribute to
    better outcomes \cite{www-google-elec}.

passed: False
```

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 15 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 [38]. This type of technology, with IoT being short for "Internet of Things", is broadly defined in its application, which could come in the form of a phone, vehicle, a home device like a thermostat or television, but rather specific in its intention. Put simply, IoT technology is intended to describe devices that collect and relay information via the Internet. They are different than a computer though in that an IoT device is built to serve a specific purpose or job, but not do the actual heavy computing itself. Nevertheless, computing developments have been central to the recent growth in the IoT. Innovations in computing power and speed have increased the amount of data that can be collected and processed, typically referred to "Big Data", which has enabled IoT devices to become more personalized and useful. The utility of these devices has spurred tremendous growth recently, on the order of 30% annually and 2017 is expected to be the year that the number of IoT devices exceeds the number of people on Earth [15].

As the growth in IoT, aided by Big Data, continues in coming years, scalability of the devices becomes a central concern. Further growth from this point means multiple IoT devices per person, which means an increasing number of providers will need to find a solution to this problem.

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 relationship between IoT and Big Data and then later 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, recent popularity of IoT, it might be a surprise 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, [13]. 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 [46].

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, which will be useful later when discussing the sample project undertaken and how these both relate to Big Data analysis. 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 more well known types, such as Wi-Fi, Bluetooth or RFID, to the perhaps less widely known, such as NFC (Near Field Communication) or Zigbee [39]. With this connection to the internet, by virtue of the amount of data they're able to now collect, make them tremendously influential in the advancement of Big Data. As will hopefully become through the progression of this paper, IoT allows its users to quantify the world around them. 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 [25]. 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 [48]. 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 [14].

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. Specifically, the Raspberry Pi comes as a truly blank slate and to begin using it, one will need to write the OS onto the Pi. Once this is completed though, the interface will display as the typical desktop that most know well. In the interest of length and since there exists several tutorials in how to initialize the Pi, the discussion below will assume that the user has been able to successfully get the Pi operational and to the Linux prompt and ultimately, Python.

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 the 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. As is likely evident by the description of the project and will be expanded on further, this application benefits greatly from the creativity by the providers in solving difficult Big Data issues, such as the clustering of news (Google) and weather forecasting (Yahoo/Weather Underground).

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 [23]. Additionally, 44% of users were found to read nothing more than the headlines [19]. 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 [10].

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 [56]. The current valuation is undoubtedly higher. 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, Axel Springer, Germany's largest news source decided to remove themselves from the index for two weeks and the results were devastating for the site as traffic through the site dropped by 40% [55].

3.6.2 Big Data Solution. While Google News has undergone many innovations, especially in recent years with demand by users to auto-detect and filter out fake news, the information that has been released is relevant, but dated. Of what has been uncovered, the primary algorithms in the early versions used in this Big Data problem were MinHash, Probabilistic Latent Semantic Indexing and Covisitation. 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 [51].

3.7 Weather Underground via Yahoo API

3.7.1 Big Data Description. Weather is a primary concern in planning for many businesses and 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, forecasts provided by the National Weather Service (NWS), which aggregates data from airports and weather balloons, are pooled along with data from personalized weather stations that are maintained by its users, which number over 250,000. This 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) [50]. 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 [21].

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 [43]. 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 Future Considerations

It is rather commonplace knowledge, for better or worse, that the apps we use daily are collecting data on us. This is one of the debates around IoT that one of its main 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" [33]. These risky behaviors, which we will explore later, are defined as activities such as tracking locations, accessing users' contact lists or selling registration data to ad agencies.

In subsequent builds of this application, if the intention was distribute to a wider audience, collecting volunteered user data would be an interesting addition. Then, this data can be pooled together for how the application could be tailored to meet geographic or demographic preferences.

Additionally, as with most apps, 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 [17].

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, individual 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 can become 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 bandwith 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 [47].

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 [34]. 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 [26].

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 a near necessity to effectively manage the problem. This is because the approval process for a transaction needs to be virtually instantaneous, but 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. [53] 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 [54]. 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 [44].

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 [12]. With IoT, this adds a tremendous amount of data to an already Big Data problem. As 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 [22]. 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 [28].

4.2 Use Case: Autonomous Vehicles

In many ways, autonomous vehicles represent the pinnacle of edge computing to date. 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 [27]. 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 [20]. 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 useful as never making the calculation in the first place, so timing is of the essence. 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 [18]. 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. This is another instance of the familiar union between IoT and Big Data. 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* [35] [16].

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 elder;y portion of the population was just 10%, but accounted for 34% of medical expenditures [31].

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.

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]. Outside of the elderly though, 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 [49]. 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 [11]. 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 somehow with that type 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 style. The 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 [30]. To compete with this personalization, 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 [29]. In fact, America's largest retailer, Wal-Mart, is currently using facial technology to sense customer's moods and find those who are dissatisfied [37].

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 [42].

Outside of this custom shopping, facial recognition is also being used to deal with a risk that e-commerce is not exposed to - shoplifting. With this technology, one retail store was able to identify when a 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.

5 CONCERNS 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 [36].

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 [41].

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 [40].

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 [32].

5.2 Privacy

Naturally, one of the consequences of a security breach via an IoT device would be having personal information comprised. 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 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 [45]. 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 [52].

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 have 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 [24].

6 CONCLUSION

As was hopefully made clear, the 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 packages required to compile the sample IoT monitor described can be found at the bottom of the following address in the README file along with a sample output. https://github.com/bigdata-i523/hid334/tree/master/project

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-cant-start-without-edge-computing

[2] Taylor Armerding. 2017. The 16 biggest data breaches of the 21st century. (10 2017). https://www.csoonline.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/#5e01a7965c8f

[9] Gary Davis. 2017. A Cybersecurity Carol: Key Takeaways From This Yearfis Most Hackable Holiday Gifts. Website. (11 2017). https://securingtomorrow.mcafee.com/consumer/consumer-threat-notices/most-hackable-gifts/

[10] 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

[11] 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

[12] 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/

[13] Arik Gabbai. 2015. Kevin Ashton Describes fithe Internet of Thingsfi. Magazine. (01 2015). https://www.smithsonianmag.com/innovation/kevin-ashton-describes-the-internet-of-things-180953749/

[14] Garmin. 2017. Garmin Edge 500. Website. (2017). https://buy.garmin.com/en-US/US/p/36728#overview

[15] Gartner. 2017. Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016. (02 2017).

[16] 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

[17] Google. 2017. Mobile App Reporting in Google Analytics - iOS. Website. (2017). https://developers.google.com/analytics/devguides/collection/firebase/ios/#how_does_it_work

[18] 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

[19] Patrick Hoge. 2010. Survey: 44% stop at Google News headlines. Website. (01 2010). https://www.bizjournals.com/sanfrancisco/stories/2010/01/18/daily24.html

[20] Nabeel Hyatt. 2017. Autonomous driving is here, and itfis going to change everything. Website. (04 2017). https://www.recode.net/2017/4/19/15364608/autonomous-self-driving-cars-impact-disruption-society-mobility

[21] IBM. 2015. IBM Plans to Acquire The Weather Companyfis 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

[22] Ajit Jaokar. 2017. Behavioural Biometrics, IoT and AI. Website. (10 2017). https://www.datasciencecentral.com/profiles/blogs/behavioural-biometrics-iot-and-ai

[23] 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/

[24] 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

[25] 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/

[26] John Kiernan. 2017. Credit Card & Debit Card Fraud Statistics. Website. (02 2017). https://wallethub.com/edu/credit-debit-card-fraud-statistics/25725/

[27] 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

[28] 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/

[29] 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

[30] 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/#12f2f9cb6d55

[31] 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

[32] 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

[33] 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/

[34] 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/

[35] 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-dataday.html

[36] 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

[37] 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/

[38] 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/

[39] 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

[40] 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

[41] Freddie Roberts. 2017. Half of US companies hit by IoT security breaches, says survey. (06 2017). https://internetofbusiness.com/half-us-iot-security-breach/

[42] Brenda Salinas. 2013. High-End Stores Use Facial Recognition Tools To Spot VIPs. Website. (07 2013).

[43] 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/

[44] 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/

[45] 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/

[46] Mark Sullivan. 2012. A brief history of GPS. Website. (08 2012). https://www.pcworld.com/article/2000276/a-brief-history-of-gps.html

[47] 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

[48] 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/

[49] Heather Thompson. 2017. Edge computing: Itfis what healthcare IoT craves. Website. (03 2017). http://www.medicaldesignandoutsourcing.com/edge-computing-healthcare-iot-craves/

[50] Weather Underground. 2017. Weather Underground - About Our Data. Website. (2017). https://www.wunderground.com/about/data

[51] Jack Vaughan. 2013. Google's big data infrastructure: Don't try this at home? Website. (10 2013). http://searchdatamanagement.techtarget.com/opinion/Googles-big-data-infrastructure-Dont-try-this-at-home

[52] Steven J. Vaughan-Nichols. 2017. fiQuHow to keep your smart TV from spying on you. Website. (03 2017). http://www.zdnet.com/article/how-to-keep-your-smart-tv-from-spying-on-you/

[53] Visa. 2017. Visa Ready and IoT Payments. Website. (2017). https://usa.visa.com/visa-everywhere/innovation/visa-ready-and-iot-payments.html

[54] Visa. 2017. Visa Ready: Biometrics. Website. (2017). https://visaready.visa.com/Biometric_program_detail.html
[55] Harro Ten Wolde and Eric Auchard. 2014. Germany's top publisher bows to Google in news licensing row. Website. (11 2014). https://www.reuters.com/article/us-google-axel-sprngr/germanys-top-publisher-bows-to-google-in-news-licensing-row-idUSKBN0IP1YT20141105
[56] Tim Worstall. 2014. If Google News Is Worth $100 Million Then Why Can't Google Pay The Newspaper Publishers? Website. (12 2014). https://www.forbes.com/sites/timworstall/2014/12/14/if-google-news-is-worth-100-million-then-why-cant-google-pay-the-newspaper-publishers/#7496b2b555a1

```
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
===========================================

        [2017-12-04 22.02.26] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 1.0s.

================================================================================
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 6420 report.tex
wc 334 project 8 7032 report.pdf
wc 334 project 8 1783 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 cahnge the number to a smaller fraction


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



bibtex
--------------------------------------------------------------------------------



label errors


bibtex errors

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

bibtex_empty_fields
--------------------------------------------------------------------------------
```

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


find ""
--------------------------------------------------------------------------------

passed: True

ascii
--------------------------------------------------------------------------------



================================================================================
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



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 "NPR" isn't style-file defined
--line 69 of file 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 7 warnings)

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


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


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

Big Data for Edge Computing

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Gregor von Laszewski
Indiana University
Smith Research Center
Bloomington, IN 47408, USA
laszewski@gmail.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

Big Data, Edge Computing i523

1 INTRODUCTION

Put here an introduction about your topic. We just need one sample refernce so the paper compiles in LaTeX so we put it here [?].

2 FIGURES

In Figure 1 we show a fly. Please note that because we use just columwidth that the size of the figure will change to the columnwidth of the paper once we change the layout to final. CHnaging the layout to final should not be done by you. All figures will be listed at the end. Please do not use phrases such as as shown in the figure below.

[Figure 1 about here.]

When copying the example, please do not check in the images from the examples into your images directory as you will not need them for your paper. Instead use images that you like to include. If you do not have any images, do not dreate the images folder.

3 TABLES

In case you need to create tables, you can do this with online tools (if you do not mind sharing your data) such as https://www.tablesgenerator.com/ or other such tools (please google for them). They even allow you to manage tables as CSV.

or generate them by hand while using the provided template in Table??. Not ethat the caption is before the tabular environment.

[Table 1 about here.]

4 QUOTES

Do not use ”these quotes” but use these “these quotes”.

5 LABELS

Do not use Figure 1 user the ref for the figure while using its label

6 FOOTNOTES

Footnotes must be avoided in papers. All URLs must be included as full refernces/citations and used with the `\cite` command [1].

[1] do not use footnotes

7 LONG EXAMPLE

If you like to see a more elaborate example, please look at report-long.tex.

8 CONCLUSION

Put here an conclusion. Conlcusions and abstracts must not have any citations in the section.

ACKNOWLEDGMENTS

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

REFERENCES

We include an appendix with common issues that we see when students submit papers. One particular important issue is not to use the underscore in bibtex labels. Sharelatex allows this, but the proceedings script we have does not allow this.

When you submit the paper you need to address each of the items in the issues.tex file and verify that you have done them. Please do this only at the end once you have finished writing the paper. To d this cange TODO with DONE. However if you check something on with DONE, but we find you actually have not executed it correcty, you will receive point deductions. Thus it is important to do this correctly and not just 5 minutes before the deadline. It is better to do a late submission than doing the check in haste.

List of Figures

1 Example caption 3

Figure 1: Example caption

List of Tables

1 My caption 5

Table 1: My caption

1	2	3
4	5	6
7	8	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
Warning--I didn't find a database entry for "editor00"
(There was 1 warning)

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


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


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


latex report
===========================================

        [2017-12-04 22.03.05] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflate
p.1   L57   : [editor00] undefined
p.1   L85   : 't:mytable' undefined
Empty `thebibliography' environment.
There were undefined citations.
bookmark level for unknown  defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
There were undefined references.
Typesetting of "report.tex" completed in 0.9s.

================================================================================
Compliance Report
================================================================================


name: Budhaditya Roy
hid:  348
paper1: 100% Oct 25 17
```

```
paper2: 100%
project: 80%

yamlcheck
--------------------------------------------------------------------------------



wordcount
--------------------------------------------------------------------------------

5
wc 348 project 5 579 content.tex
wc 348 project 5 551 report.pdf
wc 348 project 5 0 report.bib

find "
--------------------------------------------------------------------------------

103: Do not use "these quotes" but use these ``these quotes''.

passed: False

find footnote
--------------------------------------------------------------------------------

113: \footnote{do not use footnotes}.

passed: False

find input{format/i523}
--------------------------------------------------------------------------------

4: \input{format/i523}

passed: True

find input{format/final}
--------------------------------------------------------------------------------

passed: False

floats
--------------------------------------------------------------------------------

62: In Figure \ref{f:fly} we show a fly. Please note that because we
    use
```

```
69: \begin{figure}[!ht]
70: % \centering\includegraphics[width=\columnwidth]{images/fly.pdf}
71: \caption{Example caption}\label{f:fly}
86: or generate them by hand while using the provided template in
    Table\ref{t:mytable}. Not ethat
89: \begin{table}[htb]
92: \label{t:mytabble}

figures 1
tables 1
includegraphics 1
labels 2
refs 2
floats 2

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

Label/ref check
106: Do not use Figure 1 user the ref for the figure while using its
     label
passed: False -> labels or refs used wrong

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


find textwidth
--------------------------------------------------------------------------------

passed: True

below_check
--------------------------------------------------------------------------------



WARNING: figure and below may be used improperly

67: figure below.

bibtex
--------------------------------------------------------------------------------
```

```
label errors


bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--I didn't find a database entry for "editor00"
(There was 1 warning)

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
```