

# *Use Cases in Big Data Software and Analytics*

Vol. 1, Fall 2017

---

*Bloomington, Indiana*

Monday 6<sup>th</sup> November, 2017, 17:38

Editor:  
Gregor von Laszewski  
Department of Intelligent Systems  
Engineering  
Indiana University  
[laszewski@gmail.com](mailto:laszewski@gmail.com)

# Contents

<b>1 Preface</b>	<b>9</b>
1.0.1 Disclaimer . . . . .	9
1.0.2 Citation . . . . .	9
1.1 List of Papers . . . . .	10
<b>2 Biology</b>	<b>13</b>
<b>3 Business</b>	<b>13</b>
2 hid310	Status: 11/06/17 100%
Big Data Applications for Vehicle Crash Prediction	
K, e, v, i, n, , D, u, f, f, y . . . . .	13
<b>4 Edge Computing</b>	<b>24</b>
<b>5 Education</b>	<b>24</b>
<b>6 Energy</b>	<b>24</b>
3 hid224	Status: 50%
Big Data Applications in the Energy and Utilities Sector	
Rawat, Neha . . . . .	24
<b>7 Environment</b>	<b>31</b>
4 hid231	Status: 100%
Using Big Data to Battle Air Pollution	
Vegi, Karthik . . . . .	31
<b>8 Government</b>	<b>43</b>
<b>9 Health</b>	<b>43</b>
<b>10 Lifestyle</b>	<b>43</b>
5 hid332	Status: 100%
Big Data Analytics in Developing Countries	
Judy Phillips . . . . .	43
<b>11 Machine Learning</b>	<b>53</b>
<b>12 Media</b>	<b>53</b>

6 hid336	Status: 0%	
Big Data Analysis for Computer Network Defense		
Jordan Simmons . . . . .	53	
<b>13 Physics</b>		<b>59</b>
<b>14 Security</b>		<b>59</b>
7 hid316	Status: 100%	
Big Data on IoT Smart Refrigerators		
Robert Gasiewicz . . . . .	59	
8 hid329	Status: 100% Nov 6	
Big Data and the Issue of Privacy		
Ashley Miller . . . . .	70	
<b>15 Sports</b>		<b>70</b>
<b>16 Technology</b>		<b>70</b>
9 hid203	Status: 100% Nov 06 2017	
Big Data Analytics Using Regression Techniques		
Chandwani, Nisha . . . . .	70	
10 hid233	Status: Nov 10 17 80%	
Big Data Applications in Virtual Assistants		
Wang, Jiaan . . . . .	80	
11 hid302	Status: 100%	
Hadoop and MongoDB in support of Big Data Applications and Analytics		
Sushant Athaley . . . . .	89	
12 hid306	Status: 100%; 11/4/2017	
Why Deep Learning matters in IoT Data Analytics?		
Murali Cheruvu . . . . .	107	
13 hid330	Status: 100%	
MQTT for Big Data and Edge Computing		
Janaki Mudvari Khatiwada . . . . .	118	
<b>17 Text</b>		<b>118</b>
<b>18 Theory</b>		<b>118</b>
14 hid324	Status: 100% Nov 6 17	
Big Data in Decentralized election		
Ashok Kuppuraj . . . . .	118	
<b>19 Transportation</b>		<b>126</b>
<b>20 TBD</b>		<b>126</b>
15 hid101	Status: not yet started	
Benchmarking a BigData Docker deployment		
Huiyi Chen . . . . .	126	

16	hid102	Benchmarking a BigData Docker deployment Gregor von Laszewski . . . . .	Status: unkown 126
17	hid104	Big Data = Big Bias? The Fallibility of Big Data Jones, Gabriel . . . . .	Status: Nov 8 17 95% 126
18	hid105	Predictive Analytics in Sporting Match Outcomes Lipe-Melton, Josh . . . . .	Status: 100% November 6, 2017 126
19	hid106	Big Data Analytics and Insurance Fraud Detection Qiaoyi Liu . . . . .	Status: 100% 126
20	hid107	Benchmarking a BigData Docker deployment Gregor von Laszewski . . . . .	Status: unkown 134
21	hid109	Big Data and Application in Amazon Shiqi Shen . . . . .	Status: complete 100% Oct 4th 134
22	hid111	Benchmarking a BigData Docker deployment Gregor von Laszewski . . . . .	Status: unkown 141
23	hid201	Using MQTT for Communication in IoT Applications Arnav, Arnav . . . . .	Status: 80% 141
24	hid202	Visualization in Big Data. Himani Bhatt . . . . .	Status: 50% 150
25	hid204	Big Data and Support Vector Machines Chaturvedi, Dhawal . . . . .	Status: Nov 12 17 30% 150
26	hid205	This is my paper about the other abc Chaudhary Mrunal L . . . . .	Status: 0% 150
27	hid208	Algorithms for Big Data Analysis Jyothi Pranavi Devineni . . . . .	Status: Nov 5 17 100% 150
28	hid211	Machine learning optimizations for big data Khamkar, Ajinkya . . . . .	Status: unkown 154
29	hid212	Not yet decided Kumar, Saurabh . . . . .	Status: unkown 163
30	hid213	Big Data and Face Identification Yuchen Liu . . . . .	Status: Nov 06 2017 100% 163

31 hid214		Status: 100%
	Big Data and League of Legend	
	Junjie Lu . . . . .	163
32 hid215		Status: Nov 6 17 100%
	Big Data and Artificial Intelligence with Computer Vision	
	Mallala, Bharat . . . . .	173
33 hid216		Status: not started
	n/a	
	Millard, Mathew . . . . .	173
34 hid218		Status: 100%
	How Big Data Transform Education	
	Niu, Geng . . . . .	173
35 hid219		Status: unkown
	Benchmarking a BigData Docker deployment	
	Gregor von Laszewski . . . . .	182
36 hid225		Status: not started
	...	
	Schwartz, Matthew . . . . .	182
37 hid228		Status: Nov 06 17 100%
	Big Data Applications in Aviation Industry	
	Swargam, Prashanth . . . . .	182
38 hid230		Status: unkown
	Big data with natural language processing	
	YuanMing Huang . . . . .	188
39 hid232		Status: 100%
	Big Data in Rain water harvesting	
	Rahul Velayutham . . . . .	188
40 hid234		Status: Nov 8 2017 70%
	Big Data and Cloud Computing in Health Informatics for People with Disabilities.	
	Weixuan Wang . . . . .	197
41 hid235		Status: 0%
	Big data: An Opportunity for Historians	
	Yujie Wu . . . . .	203
42 hid236		Status: not started
	Benchmarking a BigData Docker deployment	
	Weipeng Yang . . . . .	209
43 hid237		Status: 0%
	Big Data Analytics in Social Media Threat Research	
	Tousif Ahmed . . . . .	209
44 hid301		Status: 100% Nov 4
	Prediction of psychological traits based on Big Data classification of associated social media footprints	
	Gagan Arora . . . . .	220
45 hid304		Status: 100%
	Big Data in Deep Space Telemetry and Navigation	
	Ricky Carmickle . . . . .	235

46 hid305		Status: 100%
	Big Data Security and Privacy.	
	Andres Castro Benavides, Uma Kugan . . . . .	235
47 hid308		Status: 0%
	Parallel Computing and Big Data	
	Pravin Deshmukh . . . . .	235
48 hid311		Status: 0%
	Benchmarking a BigData Docker deployment	
	Gregor von Laszewski . . . . .	235
49 hid312		Status: 100%
	Big Data Applications in Historical Studies	
	Neil Eliason . . . . .	235
50 hid313		Status: 100%
	Big Data Applications in Laboratories	
	Tiffany Fabianac . . . . .	244
51 hid314		Status: 0%
	Benchmarking a BigData Docker deployment	
	Gregor von Laszewski . . . . .	244
52 hid315		Status: 100% complete
	Concussions and Big Data's Opportunities and Challenges	
	Garner, Jeffry . . . . .	244
53 hid318		Status: 0%
	Benchmarking a BigData Docker deployment	
	Gregor von Laszewski . . . . .	244
54 hid319		Status: 0%
	Mini Project: ESP8266 and Raspberry PI Robot Car	
	Mani Kumar Kagita . . . . .	244
55 hid320		Status: 50%
	Overview of Python Data Visualization Tools	
	Elena Kirzhner . . . . .	244
56 hid321		Status: unkown
	Benchmarking a BigData Docker deployment	
	Gregor von Laszewski . . . . .	244
57 hid323		Status: Review Date 11.06.2017
	Big Data Security and Privacy	
	Uma M Kugan . . . . .	244
58 hid326		Status: unkown
	Big data on autonomous cars	
	Mohan Mahendrakar . . . . .	244
59 hid328		Status: 100%
	Big data analytics in data center network monitoring	
	Dhanya Mathew . . . . .	244
60 hid331		Status: Nov 6 17 100%
	Big Data Applications in Using Neural Networks for Medical Image Analysis	
	Tyler Peterson . . . . .	258

61	hid333	Natural Language Processing (NLP) to Analyze Human Speech Data Ashok Reddy Singam, Anil Ravi	Status: 100% 268
62	hid334	Advancements in Drone Technology for the US Military Peter Russell	Status: 100% 268
63	hid335	Big Health Data from Wearable Electronic Sensors (WES) and the Treatment of Opioid Addiction Sean M. Shiverick	Status: 100% 275
64	hid337	Natural Language Processing (NLP) to Analyze Human Speech Data Ashok Reddy Singam, Anil Ravi	Status: Nov 06 17 100% 297
65	hid338	A comparative study of Kubernetes and Docker Swarm and Advantages of Singularity Container to HPC World Anand Sriramulu	Status: 0% 308
66	hid339	Benchmarking a BigData Docker deployment Hady Sylla	Status: Nov 5 2017 100% 308
67	hid340	BigchainDB: A Big Database for the Blockchain? Timothy A. Thompson	Status: 0% 308
68	hid341	This is my paper about the other abc Tibenkana, Jacob	Status: 0% 308
69	hid342	Applications of Big Data Analytics in Public Policy Development and Evaluation Udoyen, Nsikan	Status: 70% 308
70	hid346	This is my paper about the other abc Gregor von Laszewski	Status: unkown 308
71	hid348	Security aspect of NOSQL database in Big Data Applications Budhaditya Roy	Status: 100% 308



# 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 regards 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	Benchmarking a BigData Docker deployment
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
104	Jones, Gabriel	Big Data = Big Bias? The Fallibility of Big Data
105	Lipe-Melton, Josh	Predictive Analytics in Sporting Match Outcomes
106	Qiaoyi Liu	Big Data Analytics and Insurance Fraud Detection
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
109	Shiqi Shen	Big Data and Application in Amazon
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
201	Arnav, Arnav	Using MQTT for Communication in IoT Applications
202	Himani Bhatt	Visualization in Big Data.
203	Chandwani, Nisha	Big Data Analytics Using Regression Techniques
204	Chaturvedi, Dhawal	Big Data and Support Vector Machines
205	Chaudhary Mrunal L	This is my paper about the other abc
208	Jyothi Pranavi Devineni	Algorithms for Big Data Analysis
209	Han, Wenxuan	Clustering Algorithms in Big Data Analysis
hid210	error: yaml	Clustering Algorithms in Big Data Analysis
211	Khamkar, Ajinkya	Machine learning optimizations for big data
212,	Kumar, Saurabh	Not yet decided
213	Yuchen Liu	Big Data and Face Identification
214	Junjie Lu	Big Data and League of Legend
215	Mallala, Bharat	Big Data and Artificial Intelligence with Computer Vision
216	Millard, Mathew	n/a
218	Niu, Geng	How Big Data Transform Education
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
224	Rawat, Neha	Big Data Applications in the Energy and Utilities Sector
225	Schwartz, Matthew	...
228	Swargam, Prashanth	Big Data Applications in Aviation Industry
229	ZhiCheng Zhu	Big Data Analytics in Mobile device Application Development
230	YuanMing Huang	Big data with natural language processing

231	Vegi, Karthik	Using Big Data to Battle Air Pollution
232	Rahul Velayutham	Big Data in Rain water harvesting
233	Wang, Jiaan	Big Data Applications in Virtual Assistants
234	Weixuan Wang	Big Data and Cloud Computing in Health Informatics for People with Disabilities.
235	Yujie Wu	Big data: An Opportunity for Historians
236	Weipeng Yang	Benchmarking a BigData Docker deployment
237	Tousif Ahmed	Big Data Analytics in Social Media Threat Research
301	Gagan Arora	Prediction of psychological traits based on Big Data classification of associated social media footprints
302	Sushant Athaley	Hadoop and MongoDB in support of Big Data Applications and Analytics
304	Ricky Carmickle	Big Data in Deep Space Telemetry and Navigation
305,	Andres Castro Benavides, Uma	Big Data Security and Privacy.
323	Kugan	
306	Murali Cheruvu	Why Deep Learning matters in IoT Data Analytics?
308	Pravin Deshmukh	Parallel Computing and Big Data
hid309	error: yaml	Parallel Computing and Big Data
hid310	error: yaml	Big Data Applications for Vehicle Crash Prediction
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
312	Neil Eliason	Big Data Applications in Historical Studies
313	Tiffany Fabianac	Big Data Applications in Laboratories
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
315	Garner, Jeffry	Concussions and Big Data's Opportunities and Challenges
316	Robert Gasiewicz	Big Data on IoT Smart Refrigerators
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
319	Mani Kumar Kagita	Mini Project: ESP8266 and Raspberry PI Robot Car
320	Elena Kirzhner	Overview of Python Data Visualization Tools
0	Gregor von Laszewski	Benchmarking a BigData Docker deployment
323	Uma M Kugan	Big Data Security and Privacy
324	Ashok Kuppuraj	Big Data in Decentralized election
325	J. Robert Langlois	The importance of data sharing and the replication of the sciences
326	Mohan Mahendrakar	Big data on autonomous cars
327	Paul Marks	The Impact of Self-Driving Cars on the Economy
328	Dhanya Mathew	Big data analytics in data center network monitoring
329	Ashley Miller	Big Data and the Issue of Privacy
330	Janaki Mudvari Khatiwada	MQTT for Big Data and Edge Computing
331	Tyler Peterson	Big Data Applications in Using Neural Networks for Medical Image Analysis
332	Judy Phillips	Big Data Analytics in Developing Countries
337,	Ashok Reddy Singam, Anil Ravi	Natural Language Processing (NLP) to Analyze Human Speech Data
333		
334	Peter Russell	Advancements in Drone Technology for the US Military
335	Sean M. Shiverick	Big Health Data from Wearable Electronic Sensors (WES) and the Treatment of Opioid Addiction
336	Jordan Simmons	Big Data Analysis for Computer Network Defense
337,	Ashok Reddy Singam, Anil Ravi	Natural Language Processing (NLP) to Analyze Human Speech Data
333		
338	Anand Sriramulu	A comparative study of Kubernetes and Docker Swarm and Advantages of Singularity Container to HPC World
339	Hady Sylla	Benchmarking a BigData Docker deployment
340	Timothy A. Thompson	BigchainDB: A Big Database for the Blockchain?

341	Tibenkana, Jacob	This is my paper about the other abc
342	Udoyen, Nsikan	Applications of Big Data Analytics in Public Policy Development and Evaluation
343	Borga Edionse Usifo	Big Data Applications and Manufacturing
hid345	error: yaml	How the Datafication of Activity is Improving Human Health
0	Gregor von Laszewski	This is my paper about the other abc
347	Jeramy Townsley	Sociological Qualitative Methods Using Big Data
348	Budhaditya Roy	Security aspect of NOSQL database in Big Data Applications

# Big Data Applications for Vehicle Crash Prediction

Kevin Duffy  
Indiana University  
4014 E. Stop 10 Rd.  
Indianapolis, Indiana 46237  
kevduffy@iu.edu

## ABSTRACT

The idea of predicting car crashes used to be a fruitless endeavor - relegated to mere guesswork. However, advances in big data applications have allowed law enforcement to more accurately predict when and where car crashes are likely to occur, thus lowering first response times and taking proactive actions to prevent accidents in high-risk corridors. This paper shows several different approaches different agencies have taken using the power of data to solve this critical problem.

## KEYWORDS

Big Data, Vehicles, Crashes, HID310, I523

## 1 INTRODUCTION

Car crashes are the top cause of death for Americans between ages 5 and 34.[8] Although the rate of car crashes has been dropping steadily over the last few decades [2], they remain a constant hazard for any who travel on our roads.

While most of the talk regarding car crashes and big data usually involves the development of self-driving cars (and understandably so), more immediate measures are being developed to meet this problem. A common usage of data in this domain is the prediction of high-risk crash areas.

The idea of predicting when and where crashes would occur used to be pure guesswork on the part of experienced police troopers[5], but advances in big data analysis has allowed law enforcement agencies to begin predicting and mapping exactly when and where high-risk areas will occur. Advocates contend that these tools allow first responders to act more efficiently both reactively and proactively, making our roads safer on which to travel.

We will outline the initial problem being addressed by these tools, the ultimate goal of such an exercise, and briefly explore two approaches to this solution. We will determine what outcomes these applications had, as well as whether they can be fully evaluated at this time.

## 2 TRENDS AND GOALS

In 1975, the United States had a rate of 3.35 deaths per 100 million miles traveled.[2] In 2015, that rate stood at 1.13 deaths. In fact, during that time frame, the amount of miles traveled on our roads has more than doubled, while the average annual number of fatalities on our roads has dropped by more than 10,000. The National Highway Traffic Safety Administration (NHTSA) credits this to the successful implementation of the Four Es of traffic safety:

- Enforcement
- Engineering
- Education

- Emergency Medical Services [8]

However, the NHTSA and other safety groups are not satisfied with merely lower numbers. They are driven by an initiative called Toward Zero Deaths which, as the name implies, strives toward the elimination of traffic related deaths.[8]

But how can highway safety groups reach this lofty goal? And what are they already doing about it? It should come as little surprise that the powerful utilization of big data has been rising in this field.

## 3 BIG DATA SOLUTIONS

A lot of data points can be collected from current activities undertaken through the Four Es. These include things such as traffic volume at the time of the crash, vehicle speed, road and construction conditions, and emergency response times. [8]

In addition, state agencies have found that other extraneous factors contributed to the likelihood of a crash, such as events like football games, major holidays that correspond with an increase in drunk driving (such as New Year's Eve and Super Bowl Sunday), and concentrations of establishments that serve alcohol.[5]

Unfortunately, these data points were previously siloed off in different agencies and formats, making dynamic usage of this information difficult if not impossible. However, some states are beginning to utilize this data in an effective way. They are beginning to collect data in a deliberate way so as to coordinate information between agencies in order to create applications and tools for use by troopers, other emergency response, road planners, and the public at large. [5]

States are going about this problem in different ways. This will allow us to approach the issue from different angles and see which approaches yield superior results. However, since these applications are still so new, it is not yet possible to fully evaluate their effectiveness on outcomes. This paper is examining two state approaches: Tennessee, which was one of the first initiatives in the nation, and Indiana, which was modeled after Tennessee's program in a more comprehensive way.

### 3.1 Tennessee

The origin of Tennessee Highway Patrol's Predictive Analytics program began in 2008, when state troopers switched from paper-and-pen crash reports to an electronic interface. This allowed Tennessee's agencies to use real-time data to create prediction software for car crashes. According to the State of Tennessee, the program uses SPSS software to apply three different statistical models with machine-learning algorithms in order to provide troopers with risk areas for certain types of crashes. [4] The models used were:

- (1) Crash Reduction Analyzing Statistical History (CRASH). Using risk factors and historical crash data, this model gives probabilities on a fatal or injury-causing accident over four-hour blocks over a one week period, which is then illustrated on interactive maps.
- (2) Driving Under the Influence (DUI). Calculates the probability of a crash related to a DUI from 4 pm to 4 am.
- (3) Commercial Motor Vehicle (CMV). Calculates the probability of a crash related to a commercial motor vehicle (such as a semi truck).[4]

The program was implemented in 2013. Using this technology, troopers stationed themselves in more statistically advantageous positions in order to decrease response times to crashes, and take preliminary actions to prevent crashes such as setting up traffic direction or more aggressive enforcement on speeding and reckless driving.

From 2013 to 2015, Tennessee traffic deaths fell 3 percent, compared to a 7 percent increase across the nation. This cannot yet be directly tied to the performance of the program, but according to officials such as the THP Statistics official Patrick Dolan, the effect is clear.

Information coming out of the Predictive Analytics program driving targeted enforcement "has allowed us to deploy our resources more effectively to execute our mission successfully," Dolan reported to the Tennessee government's Traffic Safety Innovations 2016 newsletter. [4]

However, the trend became murkier in 2016. Tennessee traffic deaths spiked 8 percent, matching the national trend. Officials are still unclear whether the model is to blame. According to Dolan, average police response time dropped nearly 33 percent between 2012 and 2016. [5]

### 3.2 Indiana

In March 2014, then-Indiana Governor Mike Pence create by executive order the Management Performance Hub (MPH), a subagency tasked with driving "a coordinated effort by state agencies to share data and improve and strengthen services, maximize the utilization of available resources, and ensure that state services are available to all Hoosiers."<sup>[3]</sup>

One of the first projects undertaken by MPH was the Crash Prediction Website. Inspired by Tennessee's program[7], the MPH worked in tandem with the Indiana State Police to create an interactive map (Figure 1) showing the probability of both fatal and nonfatal traffic accidents.

[Figure 1 about here.]

The model grew out of several factors:

- (1) The concept was borrowed from Tennessee's aforementioned Predictive Analytics program, though Indiana had a greater range and depth of data available.
- (2) The core of the model is built upon data from 2 million crashes going back to 2004. The data was cleaned so only relevant crash data was included in the forecast.
- (3) The probability of a car crash is ranked from "very low risk" to "high risk", which is then indicated on the map through color-coded 1 square kilometer blocks over three hour time blocks.[6] This is more granular than Tennessee's

maps, in both time and space scale. These probabilities are ranked based off of weather forecasts, traffic congestion, road conditions, time of day, historical information, and census data.

- (4) The map is populated with different colored dots to represent past car crashes - red for injury-causing, grey for non-injuries. Users can then click these to identify unique details of actual car crashes.[7]
- (5) The model is updated automatically with fresh crash reports, providing the software with dynamic information.

Unlike Tennessee's program, Indiana's crash map is completely available to the public for their own personal use, though its usage will primarily be used by police and first responders.

Although it is still too soon after launch to evaluate the effectiveness of this model, Indiana officials are optimistic. Indiana's Office of Management and Budget, the parent agency to the MPH, estimates that even a one percent reduction in Indiana crashes could net up to \$30 million in annual savings, besides the obvious benefits of fewer road casualties.

With the groundwork laid in the crash map, officials are hoping to utilize the technology in other ways. Major Michael White of the Indiana State Police, in an interview with Statescoop.com, said with the technology developed through this initiative, they hope to move on to using it to map crime data as well.[6]

## 4 CONCLUSION

While these applications and initiatives are still too new to fully evaluate, there does appear to be preliminary results that show promise.

Tennessee showed quick decline in car crashes and police response times, while Indiana built upon Tennessee's example to provide a more comprehensive look at all risk factors involved in a car crash, while opening this tool up to the public for personal use.

These applications also show promise in application to other domains, such as the Indiana State Police's interest in creating a crime risk map using the same principles.

It is also encouraging to view these as an example in various state agencies coordinating in order to share data and collaborate on applications. If Indiana's Management Performance Hub is any indication, these collaborations will continue, at least in Indiana. Hopefully more states follow suit in creating data sharing hubs and protocols in order to streamline government data utilization.

Ultimately, what is primarily needed is more time to evaluate the effectiveness of these initiatives, as well as metadata related to utilization of the map (for example, what is the average response times of a trooper using the map versus one who is not using the map). As more states grow their own initiatives, we can also evaluate whether certain approaches are more effective than others.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and all TA's for their tireless work in ensuring this class goes smoothly.

## REFERENCES

- [1] [n. d.]. ([n. d.]). <https://www.in.gov/isp/ispCrashApp/main.html>
- [2] [n. d.]. General statistics. ([n. d.]). <http://www.iihs.org/iihs/topics/t-general-statistics/fatalityfacts/overview-of-fatality-facts>

- [3] 2014. . Number 14-06. 2 pages.
- [4] 2016. *State of Tennessee* (2016). <https://www.tn.gov/assets/entities/safety/attachments/Technology.v1.pdf>
- [5] Jenni Bergal. 2017. Troopers Use 'Big Data' to Predict Crash Sites. (Feb 2017). <http://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2017/02/09/troopers-use-big-data-to-predict-crash-sites>
- [6] Mackenzie Carmen. 2016. Indiana State Police unveil Daily Crash Prediction Map. (Nov 2016). <http://statescoop.com/indiana-state-police-unveil-car-crash-forecast-map-to-help-reduce-traffic-accidents-in-indiana>
- [7] Eyragon Eidam. 2016. Indiana Launches Predictive Crash Tool for Citizens, First Responders. (Nov 2016). <http://www.govtech.com/data/Indiana-Launches-Predictive-Crash-Tool-for-Citizens-First-Responders.html>
- [8] Melissa Savage. 2012. Crash analytics: How data can help eliminate highway deaths. (Apr 2012). <https://gcn.com/articles/2012/04/20/data-analytics-traffic-safety-toward-zero-deaths.aspx>

## 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 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

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

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

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
```

## 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 float. For this class, you should place all figures at the end of the report.

LIST OF FIGURES

- 1 The map shows the probability of a crash through color-coded blocks.[1]

6

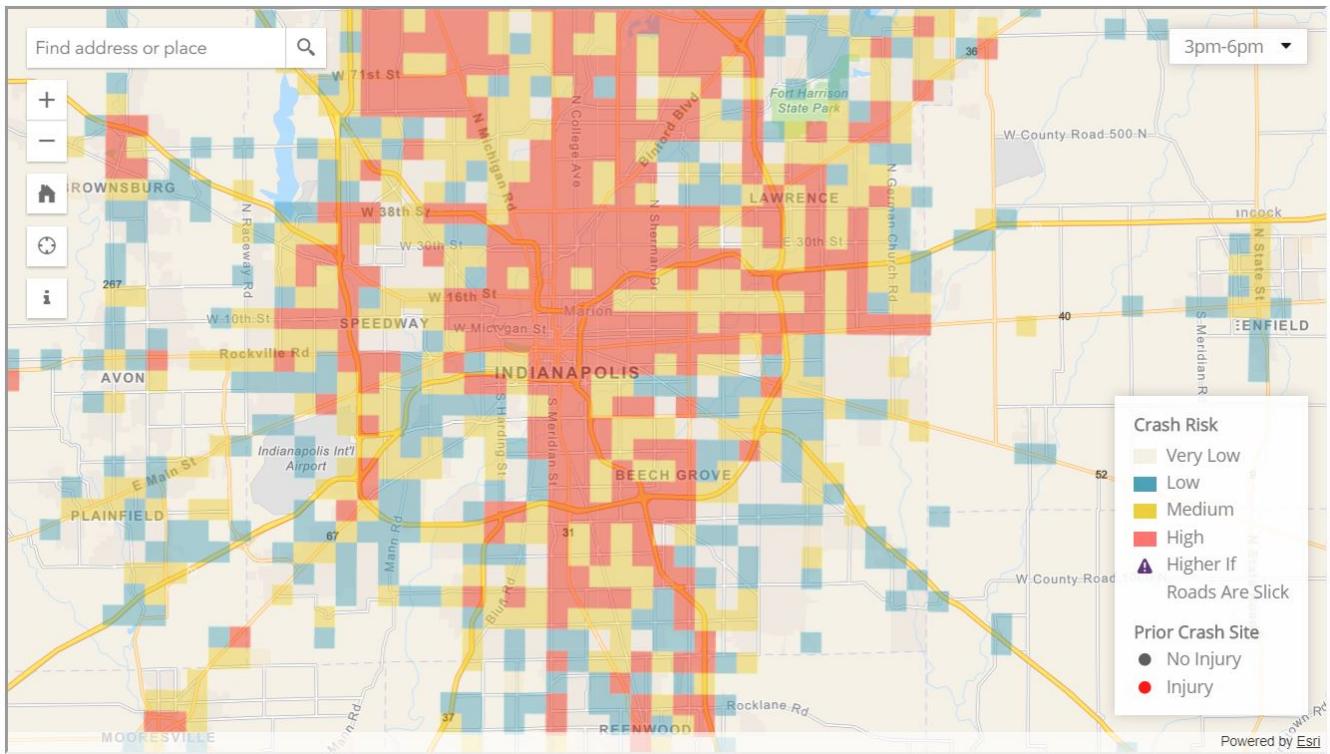


Figure 1: The map shows the probability of a crash through color-coded blocks.[1]

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--no key, author in iihs  
Warning--no author, editor, organization, or key in iihs  
Warning--to sort, need author or key in iihs  
Warning--no key, author, or editor in tennessee  
Warning--no author, editor, organization, or key in tennessee  
Warning--to sort, need author, editor, or key in tennessee  
Warning--no key, author, or editor in pence  
Warning--no author, editor, organization, or key in pence  
Warning--to sort, need author, editor, or key in pence  
Warning--no key, author in indcrashmap  
Warning--no author, editor, organization, or key in indcrashmap  
Warning--to sort, need author or key in indcrashmap  
Warning--no key, author in iihs  
Warning--no key, author in iihs  
Warning--no key, author in indcrashmap  
Warning--no key, author in indcrashmap  
Warning--no key, author, or editor in pence  
Warning--no key, author, or editor in pence  
Warning--no key, author, or editor in tennessee  
Warning--no key, author, or editor in tennessee  
Warning--no key, author in indcrashmap  
Warning--no author, editor, organization, or key in indcrashmap  
Warning--empty author in indcrashmap  
Warning--empty year in indcrashmap  
Warning--no key, author in iihs  
Warning--no author, editor, organization, or key in iihs  
Warning--empty author in iihs  
Warning--empty year in iihs  
Warning--no key, author, or editor in pence  
Warning--no author, editor, organization, or key in pence  
Warning--empty author and editor in pence  
Warning--there's a number but no series in pence  
Warning--empty publisher in pence  
Warning--empty address in pence  
Warning--no key, author, or editor in tennessee  
Warning--no author, editor, organization, or key in tennessee  
Warning--neither author and editor supplied for tennessee  
Warning--empty title in tennessee

```
Warning--no number and no volume in tennessee
Warning--page numbers missing in both pages and numpages fields in tennessee
(There were 40 warnings)
```

```
bibtext _ label error
```

```
=====
```

```
report.bib:7:@article{tennessee, url={https://www.tn.gov/assets/entities/safety/attachme
```

```
bibtext space label error
```

```
=====
```

```
report.bib:2:@book{pence, number={14-06}, year={2014}, pages={2}}
```

```
bibtext comma label error
```

```
=====
```

```
latex report
```

```
=====
```

```
[2017-11-06 17.37.17] pdflatex report.tex
```

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

```
Missing character: ""
```

```
bookmark level for unknown defaults to 0.
```

```
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
```

```
Typesetting of "report.tex" completed in 1.3s.
```

```
./README.yml
```

```
18:1      error      trailing spaces  (trailing-spaces)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Kevin Duffy
```

```
hid: 310
```

```
paper1: 11/20/17 99%
```

```
paper2: 11/06/17 100%
```

```
project: 5%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
```

```
wc 310 paper2 6 1758 report.tex  
wc 310 paper2 6 2466 report.pdf  
wc 310 paper2 6 112 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
87: \begin{figure}
```

```
88: \includegraphics[width=\columnwidth]{images/indcrashmap.jpg}
```

```
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 referred to: (refs >= labels)

Label/ref check

84: One of the first projects undertaken by MPH was the Crash Prediction Website. Inspired by Tennessee's program\cite{govtech}, the MPH worked in tandem with the Indiana State Police to create an interactive map (Figure 1) showing the probability of both fatal and nonfatal traffic accidents.

passed: False -> labels or refs used wrong

When using figures use columnwidth

[width=1.0\columnwidth]

do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

1: indcrashmap, url: do not use ' ' (spaces) in labels:  
2: pence, number: do not use ' ' (spaces) in labels:  
3: govtech, title: do not use ' ' (spaces) in labels:  
4: iihs, title: do not use ' ' (spaces) in labels:  
5: gcn, title: do not use ' ' (spaces) in labels:  
6: pew, title: do not use ' ' (spaces) in labels:  
7: tennessee, url: do not use ' ' (spaces) in labels:  
8: statescoop, title: do not use ' ' (spaces) in labels:

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

Database file #1: report.bib

Warning--no key, author in iihs

Warning--no author, editor, organization, or key in iihs  
Warning--to sort, need author or key in iihs  
Warning--no key, author, or editor in tennessee  
Warning--no author, editor, organization, or key in tennessee  
Warning--to sort, need author, editor, or key in tennessee  
Warning--no key, author, or editor in pence  
Warning--no author, editor, organization, or key in pence  
Warning--to sort, need author, editor, or key in pence  
Warning--no key, author in indcrashmap  
Warning--no author, editor, organization, or key in indcrashmap  
Warning--to sort, need author or key in indcrashmap  
Warning--no key, author in iihs  
Warning--no key, author in iihs  
Warning--no key, author in indcrashmap  
Warning--no key, author in indcrashmap  
Warning--no key, author, or editor in pence  
Warning--no key, author, or editor in pence  
Warning--no key, author, or editor in tennessee  
Warning--no key, author, or editor in tennessee  
Warning--no key, author in indcrashmap  
Warning--no author, editor, organization, or key in indcrashmap  
Warning--empty author in indcrashmap  
Warning--empty year in indcrashmap  
Warning--no key, author in iihs  
Warning--no author, editor, organization, or key in iihs  
Warning--empty author in iihs  
Warning--empty year in iihs  
Warning--no key, author, or editor in pence  
Warning--no author, editor, organization, or key in pence  
Warning--empty author and editor in pence  
Warning--there's a number but no series in pence  
Warning--empty publisher in pence  
Warning--empty address in pence  
Warning--no key, author, or editor in tennessee  
Warning--no author, editor, organization, or key in tennessee  
Warning--neither author and editor supplied for tennessee  
Warning--empty title in tennessee  
Warning--no number and no volume in tennessee  
Warning--page numbers missing in both pages and numpages fields in tennessee  
(There were 40 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 Applications in the Energy and Utilities Sector

Neha Rawat  
Indiana University  
Bloomington, Indiana  
nrawat@iu.edu

## ABSTRACT

Efficient management and utilization of energy and other utilities is the need of the hour. The plethora of real-time data generated during day-to-day operational activities can be used to detect consumption patterns and predict outages, shortages and surges in power usage, while simultaneously improving the use of renewable resources as sustainable alternatives. Intelligent big data analytics can help the energy and utilities sector by reducing costs through devising efficient operational strategies, becoming more self-sufficient and productive in their performance and improving customer satisfaction and interaction by making valuable suggestions to the consumers on how to use their resources better.

## KEYWORDS

i523, HID224, Smart Grids, Energy Disaggregation, Demand-Side Management, Sustainability, Water Management

## 1 INTRODUCTION

Energy sources such as electricity and fossil fuels like coal and petroleum, along with renewable solar and wind energy, coupled with other utilities like water and gas, are indispensable entities for humans in their day-to-day processes. We can therefore imagine the pressure on the energy and utilities sector to provide uninterrupted resource flow while ensuring efficient management of those resources. Apart from this, sustainability and use of cleaner energy is also a demand upon these industries. In earlier times, the interaction used to be a one-way street, with the industries adjusting their supply capacities in order to meet demands of the consumers. With time, the demands have increased exponentially, and the supply needs to keep pace with it. This results in issues of demand management, operational inefficiencies and increasing strain on available resources. Therefore, the energy and utilities sector too has turned to Big Data analytics for a solution. The objectives are to design intelligent systems, using the wealth of data accumulated by the energy and utilities sector, which can assist in generating, storing and using energy sustainably to meet consumer needs, while keeping costs in check (cite1). New analytics systems designed for these purposes have the capability to actively store millions of records per second from distributed sources, analyze these streams of events to detect patterns useful for prediction and constantly self-learn from previous responses using advanced cognitive capabilities (cite1). Figure 1 shows how IBM's event-driven data management system works as an efficient analytics tool for the energy sector.

The advantages that these systems can provide utilities will be visible in form of cost reductions (increasing capital productivity and saving excess expenditures on operations and maintenance), increased reliability (predicting outages and accurately detecting

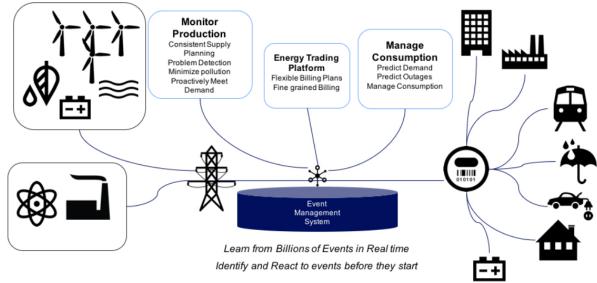


Figure 1: IBM's event-driven Data Management System (cite1)

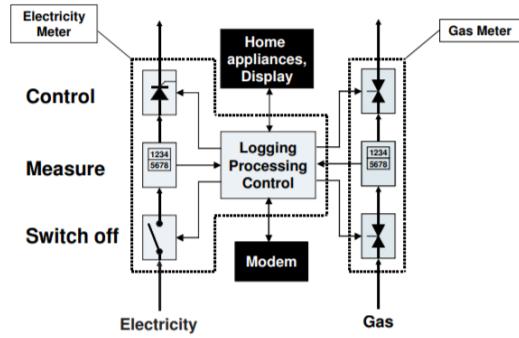
failures of equipments) and customer satisfaction (engaging customers in the process flow by providing them with useful insights about their consumption patterns) (cite2). Some of these smart technologies have been actively deployed as well. GTM Research predicts that “global utility company expenditure on data analytics will grow from 700 million in 2012 to 3.8 billion in 2020, with gas, electricity and water suppliers in all regions of the world increasing their investment” (cite3).

## 2 THE RISE OF SMART TECHNOLOGIES

Big Data warehouses and analytic technologies have been making waves in the energy and utilities sector for some time now. One example is of Microsoft, where 30,000 existing sensors were organized into a single energy-efficient system, at the company's Redmond, Washington, headquarters (cite4). The network is used to avail billions of data points on energy usage in areas such as heating, cooling and lighting. Analysis of this data lead to, in one case, a garage exhaust fan, that had been running for a year and costing Microsoft 66,000 USD. Through this system, the company saves close to a “60 million USD capital investment in energy-efficient technologies” (cite4). On a larger scale, there is huge amount of data available, being generated from oil wells, electricity and other utility grids, and generation stations. Using big data technologies coupled with the Internet of Things (IoT) i.e. smart sensors, all of this information can be gathered, structured and analyzed to provide valuable insights on utility management.

### 2.1 Smart Meters and Grids

A *Smart Meter* differs from a regular meter in its additional abilities of not only measuring the energy consumption for the customer, but also processing it and providing real-time feedback regarding it. Some of the features of a smart meter are as follows: real-time registration of energy usage, possibility to get meter information locally and remotely, remote access of the meter for adjustment of



**Figure 2: Typical Smart Meter Structure (cite5)**

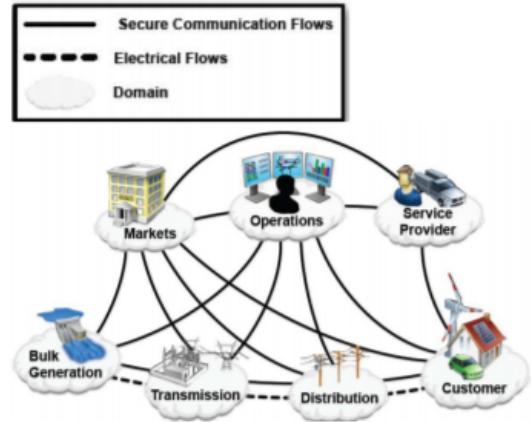
throughput, interconnection among various devices on the premise, ability to read other commodity meters in the vicinity (cite5). Figure 2 shows a typical smart meter structure.

The *smartness* of a smart meter lies in its communication system. The meter can communicate using a Power Line Carrier, a wireless modem (GSM) or an existing internet connection. An interface can be used to connect this meter to appliances and a home display, using which it can show the energy data and costs to the consumer (cite5).

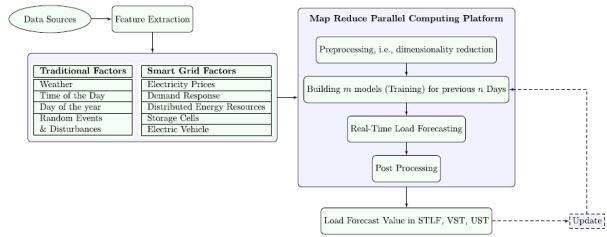
Though generally used for measuring energy consumption, smart meters can also be employed for other utilities such as gas and water. Smart water meters are not as common, but if implemented properly, can help detect issues such as leaks on the premises, in the main line, and wastage of water, much more promptly than with traditional technologies (cite9). Smart metering technologies, in general, can prove to be an essential addition to demand response as well as predictive management techniques.

A *Smart Grid* network is an advanced form of a traditional power grid (the concept is generally applied in the electricity sector). It provides a two-way exchange of electricity and information to create a widely distributed energy-delivery system, which is reliable, resilient and sustainable (cite6). Technologies such as smart meters act as components of a smart grid framework, which acts as an intelligent system that monitors generation, transmissions and consumption in the complete electric grid and performs dynamic energy management. For example, in case of a transformer failure, the smart grid would detect it and modify the power flow such that it recovers the power delivery service (cite6). Apart from this, smart grids can also be used in shaping energy demand profiles. The three major components in a smart grid system are: smart infrastructure, smart management and smart protection. The smart infrastructure helps in advanced energy flow, monitoring and communication. The smart management system provides control services. The smart protection system ensures reliability, safety and security of the network (cite6). Figure 3 shows the NIST conceptual model for a Smart Grid.

Thus, we see that the volume of data obtained from smart meters in smart grid networks, along with other components, can be used for a variety of intentions. For example, Diamantoulakis, Kapinas and Karagiannidis have used smart grid information for the purpose of load synchronization. Cloud computing technologies have been



**Figure 3: Conceptual model for a Smart Grid (cite6)**



**Figure 4: Smart Grid Forecast Model (cite7)**

used to manage the big data obtained from a smart grid (using distributed data management and parallelization). Next, dimensionality reduction has been applied to keep only the useful predictors and data mining techniques like Artificial Neural Networks or Clustering have been used to model customer load curves. This has been followed by short-term load forecasting techniques (using regression, time-series or state-space models) to provide values for price and demand forecasts (cite7). Figure 4 shows a rough structure of the Smart Grid forecast model.

## 2.2 Energy Disaggregation

*Energy Disaggregation* is a novel idea born out of the possibilities offered by the smart metering technologies discussed above. It involves the breakdown of the main electric signal into consumption by each individual appliance in a unit. Also referred to as NILM (Non-Intrusive Load Monitoring), this technology can help create itemized energy bills for consumers and help them monitor their consumption in a much more specific format. This in turn helps in efficient management of the power grid as well, as the user can detect if any device is faulty or consuming more than the normal amount of energy (cite8). The data generated by smart meters can be used for this process. The main electric signal can then be broken down into signals from individual appliances by identifying their signatures. This data obtained on an individual as well as aggregate level can be analyzed using data mining techniques such as deep learning (neural networks), Combinatorial Optimization or

Factorial Hidden Markov Models (FHMM) to detect patterns useful for prediction purposes (cite8).

### **2.3 Electric Vehicles**

The advent of *Electric vehicles* has largely reduced the strain on non-renewable fossil fuels. They form an important part of the Smart Grid network discussed in the previous section. The increase in use of electric vehicles leads to beneficial reduction in carbon emissions as well. However, one major consideration in the use of electric vehicles is charging these vehicles without overloading the grid (cite10). However, data analytics can help us here by designing scheduling systems for charging of these vehicles. Control mechanisms need to be set up to guide the charging of these vehicles. Also, consumers need to be explained to or incentivized so that they follow these guidelines and ensure that the load on the electric grid is stabilized. The advantage of electric vehicles is the presence of large batteries which can store charge and often help in load shifting within the owner's home. This indirectly provides energy back to the grid and helps stabilize it as well. Optimization techniques, using the data obtained from use of these vehicles (their charging-discharging cycles and energy demands), can help in devising such load scheduling systems which prove to be beneficial units of the Smart Grid system (cite10).

## **3 GROWTH AREAS IN THE ENERGY INDUSTRY**

The advancement of technology has led to a cultural shift in the energy industry as well. Not only are technologies changing, but also the mindsets and therefore business strategies of companies in the energy sector (cite11). Some of the major areas in the industry touched by the advent of big data technologies are operations, asset management, demand-side management and customer satisfaction. Apart from these areas, the concept of sustainability is one of utmost importance. Efficient usage of renewable resources is another major benefit provided by data analytics to the energy industry.

### **3.1 Asset Management and Operational Efficiency**

## **4 WATER MANAGEMENT**

## **5 CASE STUDIES**

## **6 CONCLUSIONS**

## **REFERENCES**

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-11-06 17.36.06] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Empty 'thebibliography' environment.
Typesetting of "report.tex" completed in 0.9s.
```

---

```
Compliance Report
```

---

```
name: Rawat, Neha
hid: 224
paper1: Nov 3 17 100%
paper2: 50%
project: In progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
3
wc 224 paper2 3 1738 content.tex
wc 224 paper2 3 1646 report.pdf
wc 224 paper2 3 513 report.bib
```

```
find "
```

---

```
31: \includegraphics[width=\columnwidth]{images/"IBM EventStore".pdf}
```

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

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

---

```
passed: False
```

```
floats
```

---

```
30: \begin{figure}
```

```
31: \includegraphics[width=\columnwidth]{images/"IBM EventStore".pdf}
```

```
33: \label{F:event}
```

```
35: Figure \ref{F:event} shows how IBM's event-driven data management  
system works as an efficient analytics tool for the energy  
sector.\\"
```

```
44: \begin{figure}
```

```
45: \includegraphics[width=\columnwidth]{images/smart_meter.pdf}
```

```
47: \label{F:smart}
```

```
49: Figure \ref{F:smart} shows a typical smart meter structure.\\"
```

```
53: \begin{figure}
```

```
54: \includegraphics[width=\columnwidth]{images/smart_grid.pdf}
```

```
56: \label{F:grid}
```

```
58: Figure \ref{F:grid} shows the NIST conceptual model for a Smart  
Grid.\\"
```

```
60: \begin{figure}
```

```
61: \includegraphics[width=\columnwidth]{images/sg_forecast.pdf}
```

```
63: \label{F:forecast}
```

```
65: Figure \ref{F:forecast} shows a rough structure of the Smart Grid  
forecast model.
```

```
figures 4
```

```
tables 0
```

```
includegraphics 4
```

```
labels 4
```

```
refs 4
```

```
floats 4
```

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

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

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

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

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
```

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

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

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

```
find newline
```

---

```
passed: True
```

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

```
find cite {
```

---

```
passed: True
```

# Using Big Data to Battle Air Pollution

Karthik Vegi

Indiana University Bloomington

2619 East 2nd Street, Apt 11

Bloomington, IN 47401, USA

kvegi@iu.com

## ABSTRACT

We have come a long way from the stone age to build large scale industries, big cities, bullet trains, and a booming automobile industry. Technological and industrial advances are making our cities smarter by the day and yet a nagging side-effect is air pollution. Air pollution is not only creating local health hazards like respiratory and heart problems, but also directly leading to an increase in temperatures and contributing to global warming. We show how the advances in *Big Data*, *Cloud Computing*, and *Internet of Devices* can be used to combat air pollution.

## KEYWORDS

i523, hid231, big data, environment, air pollution, global warming

## 1 INTRODUCTION

Air pollution is no longer a local problem. It is a global environmental issue which involves individual countries to come together and devise measures to combat it [5]. It is causing about 3.7 million premature deaths worldwide from cardiovascular and respiratory diseases and also ruins the crops that feed the world [5]. Air pollution also has a direct effect on a number of environmental issues like global warming, depletion of ozone layer, acid rains, and impacts wild-life [5].

Back in the year 1990, the job of a typical air quality scientist was to develop atmospheric dispersion models to evaluate the air pollution caused by industries and make sure that it is within the permissible level suggested by the *Environmental Protection Agency* [2]. These models gather historic data of many years from airports and weather balloons to predict the pollution with the help of meteorology theory [2]. Although the methods used to derive the values were good enough, the limitations with respect to the technology posed a real challenge which took weeks to run the simulations, only to be cut-off in the middle due to power and storage issues [2]. The data processing engine was built on Sun-Solaris workstations with tapes handling the data storage [2]. The work-stations set up in major points in the country would communicate using a very slow network connection [2]. The data processing would be done locally and later written to all the servers which would then be split and distributed among many machines and consolidated in the end [2]. “If only we had that much more data and that much more ability to handle it, we could iterate through the model at a much finer scale. Real-time data processing remained a pipe dream” [2].

## 2 AIR POLLUTION AS A BIG DATA PROBLEM

The advent of *Big Data* and the technological advances changed the way the data is ingested and analyzed [2]. The network speeds have increased, wide range of sensors are available to collect data with a lot of precision which would feed the high speed data processing systems. Batch processing has become easier with *Hadoop* and *Map-Reduce*. The storage mechanisms have become cheaper and more disaster proof.

*IBM* is helping Beijing combat air pollution by analyzing huge amounts of data using a data analysis platform *Green Horizons* [3]. *IBM* has signed up partnerships with different cities in China and India to deploy *Big Data Analytics*, *Machine Learning* and *Internet of Things* to improve traffic, keep a check on the pollution from industrial machines, and other pollution causing agents [3]. *IBM* will deploy sensors in various places to collect data in real-time and analyze previous weather forecasts, and build improved iterative models over time [3]. The system continuously streams data from the sensors and improves the forecast by learning over time using *Machine Learning* algorithms [3]. Figure 1 shows the *Green Horizons* air quality management for Beijing.

[Figure 1 about here.]

*IBM* is collaborating with the United Nations to push the use of technological advances by every country for the common good of the world [3]. More and more cities and countries are opening air quality data to public where you can get reports in real time [1]. The *BreezoMeter* is the first mobile application that provides real-time information of the street’s air quality information using geo-location maps [1]. *Copernicus* is another monitoring service that ingests data from satellites and on site sensors on land, air and sea to provide continuous information to the users [1]. *Open Data Week* is an intergovernmental organization where 34 states come together to bring reforms and discuss how to use technology and services like *Copernicus* that use *Big Data* to test prototypes of new products to ensure they operate within the permissible levels of pollution [1].

While these initiatives help bring awareness about the seriousness of the issue, each state and country should take strict measures to bring out reforms that will help eradicate pollution. *Big Data* might never replace the environmental responsibility but it will help to plan the vision for environmental awareness and its tools make it easier to achieve the vision [1]. These tools can also be used to gauge the alternative sources of energy and the feasibility of tapping into other natural resources ensuring responsible consumption of energy [1]. For example, *IBM Bluemix* analyzed data from a steel industry and the analysis uncovered an interesting insight that the furnace wastes a lot of energy to offset the temperature of the smoke which resulted in optimizing its operation [2].

### 3 BIG DATA TECHNIQUES TO COMBAT POLLUTION

#### 3.1 Random Forest Approach for predicting air quality in Urban Sensing Systems

Air pollution in an urban setting is very important to monitor because of the population density. Air quality in these areas varies a lot in various parts of the city owing to traffic and presence of industries [6]. A random forest approach ingests data from meteorology, urban sensors, road information, and real-time traffic and predicts the air quality with utmost precision [6]. Real-time air quality information consists of measuring the concentration of  $PM_{2.5}$ ,  $PM_{10}$ , and  $NO_2$  [6].

The *Air Quality Index* AQI is the measure that is used to understand how polluted the air is [6]. AQI is measured by reading the concentration of 6 pollutant gases namely, sulfur dioxide  $SO_2$ , nitrogen dioxide  $NO_2$ , air particles smaller than  $10\ \mu m$   $PM_{10}$ , air particles smaller than  $2.5\ \mu m$   $PM_{2.5}$ , carbon monoxide  $CO$ , and ozone  $O_3$  [6]. Based on the level of AQI, the air quality is classified as shown in Figure 2.

[Figure 2 about here.]

*Traffic Congestion Status* TCS, explains the traffic status at the current hour [6]. Figure 3 shows how colors are used to represent the traffic congestion [6].

[Figure 3 about here.]

#### 3.2 RAQ Algorithm

The RAQ algorithm collects data from air monitoring station AQI, meteorology data MD, traffic congestion TCS, road information RI, and point of interest POI which is the specific location that someone is interested to visit [6]. The data refresh rate is one hour and the data is collected from different parts of the city which are divided in grids from  $G_1$  to  $G_n$  [6]. The data is divided into training and testing data sets to train the model and evaluate the model [6]. Figure 4 shows the structure of the data.

[Figure 4 about here.]

A decision tree is used to split and classify the data and the results are aggregated by collecting the data from all the sub-trees [6]. Figure 5 illustrates the procedure of RAQ.

[Figure 5 about here.]

The *Random Forest* algorithm is employed using the tree type classifier to recursively partition the dataset and generate sub-trees and finally aggregate the results of each sub-tree [6]. Each sub-tree is constructed using *Bootstrap Aggregating* where each data set is divided into different buckets by using statistical samples [6]. Once the trees are constructed, each subset of data is fed into a decision tree and the estimated AQI index is calculated [6]. The final AQI index is determined as the maximum value out of all the individual values [6]. Figure 6 shows the step-by-step RAQ algorithm.

[Figure 6 about here.]

### 4 MACHINE LEARNING MODELS

*Machine Learning* deals with augmenting computers with the ability to learn from data and program themselves [4]. These algorithms can be used to evaluate the air quality [4].

#### 4.1 Artificial Neural Network Model

Artificial Neural Network Model tries to solve the problem by simulating the functioning of brain and neurons [4]. The model architecture is a function of a sigmoid [4]. For this experiment, the air quality data was divided into training, test, and validation data with split of 60, 20, and 20 with a back propagation network of two hidden layers [4]. To ensure consistency, the air quality data for the training and test sets are derived from the same season [4]. The air quality is forecast by looking at the historic data where the input and output are represented by the air quality data measured at different times [4]. The model turns out to be reliable with a good prediction accuracy with the lowest mean square error of  $3.7 \times 10^{-4}$  [4]. The Artificial Neural Network Model is combined with *Markov Chains* to develop a new improved model with improved prediction accuracy where the ANN computes the primary values and the results are re-computed and improved by the markov transitional probability matrices [4]. Figure 7 shows the Artificial Neural Network Model with two hidden layers.

[Figure 7 about here.]

#### 4.2 Least squares Support Vector Machine Model

Least squares support vector machine is a supervised learning model used for classification and regression analysis which arrives at the solution by solving the data represented in the form of linear equations [4]. For this model, the sample data was collected from 100 sensor points in different intervals of time and at different geographical locations that ranged from urban areas with population, areas near the airport, water surface areas like lakes, and sewage processing areas [4]. The sample data was a good split with 80 percent collected from urban sewage area and the other data collected from air surface areas [4]. The fluorescence content in the air was analyzed by a portable air quality measuring device developed in-house by Zhejiang University [4]. The fluorescence data captured using the device is highly dimensional and non-linear and therefore data pre-processing is essential to bring the dimensions down to a manageable level [4]. This eliminates the ambient noise and the temperature drift from the data [4]. The algorithm predicts the regression model by looking at the training data for each cluster [4]. Finally, the vector cosine distance is used to classify the sample into clusters and the performance criterion such as *Root Mean Square Error* and *Mean Absolute Error* are computed which demonstrate the efficiency of the algorithm [4]. Figure 8 shows the pictorial representation of the algorithm.

[Figure 8 about here.]

### 5 CONCLUSION

While the new age technologies have a big role to play in measuring, tracking, and keeping air pollution in check, each person should have individual environmental responsibility to make the world a better place to live in. *Internet of Things* and *Machine Learning* are augmenting the *Big Data* capabilities like never before. This ensures that we have more data points to work in a given time and continuous data streaming means more accurate real-time analytics with efficient *Machine Learning* algorithms. All these

three technologies will continue to work in tandem to keep a check on air pollution and the imminent threats.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Ferrovial Blog. 2017. Big data will control pollution in your city. Webpage. (April 2017). <http://blog.ferrovial.com/en/2017/04/big-data-pollution-control-in-cities/>
- [2] Jay Hardikar. 2017. Environmental analysis in the era of cloud and big data platforms. Webpage. (Jan. 2017). <https://www.ibm.com/blogs/bluemix/2017/01/environmental-analysis-era-cloud-big-data-platforms/>
- [3] Alexander Howard. 2015. How IBM Is Using Big Data To Battle Air Pollution In Cities. Webpage. (Sept. 2015). <https://www.ibm.com/blogs/bluemix/2017/01/environmental-analysis-era-cloud-big-data-platforms/>
- [4] Gaganjot Kaur Kang, Jerry Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. 2017. Air Quality Prediction: Big data and Machine Learning Approaches. *International Conference on Sustainable Environment and Agriculture* 1 (10 2017).
- [5] Research Applications Laboratory. 2016. Air Pollution: A Global Problem. Webpage. (April 2016). <https://ral.ucar.edu/pressroom/features/air-pollution-a-global-problem>
- [6] Ruiyun Yu, Yu Yang, Leyou Yang, Guangjie Han, and Ogutu Ann Move. 2016. RAQ: A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors* 16 (2016), 1–86. <http://www.mdpi.com/1424-8220/16/1/86>

#### LIST OF FIGURES

1	Green Horizons air quality management for Beijing [3]	5
2	AQI classification [6]	5
3	Traffic Congestion[6]	6
4	Structure of RAQ data[6]	6
5	Procedure of RAQ [6]	7
6	RAQ Algorithm [6]	7
7	Artificial Neural Network(ANN) Model [4]	8
8	Least squares Support Vector Machine Model [4]	8

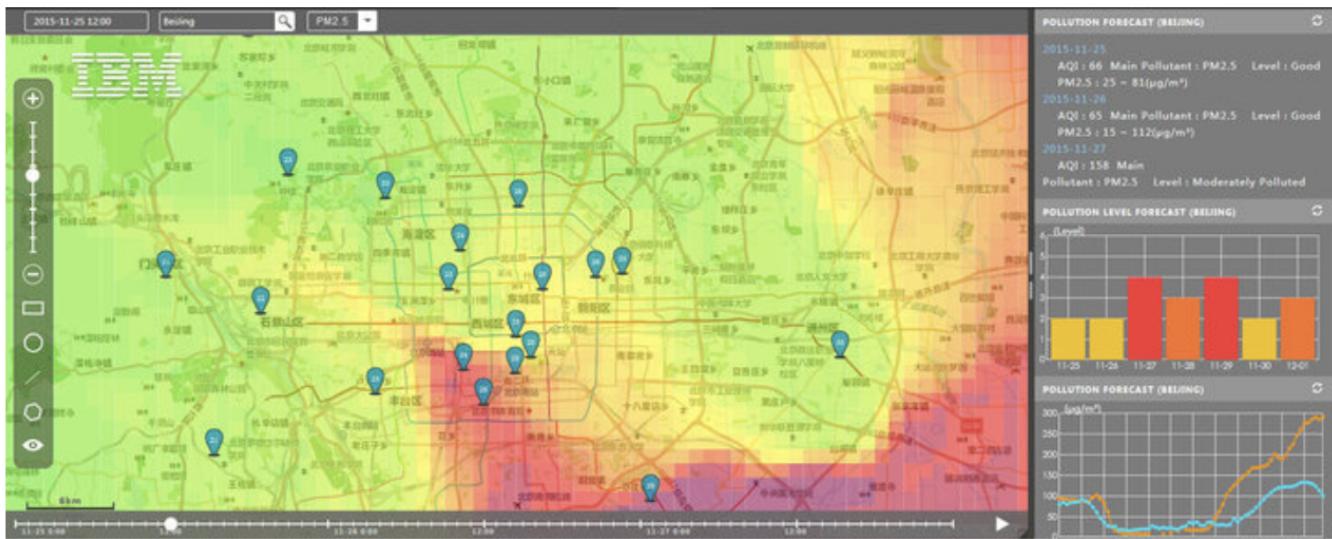


Figure 1: Green Horizons air quality management for Beijing [3]

AQI	Air Pollution Level
0–50	Excellent
51–100	Good
101–150	Lightly Polluted
151–200	Moderately Polluted
201–300	Heavily Polluted
300+	Severely Polluted

Figure 2: AQI classification [6]

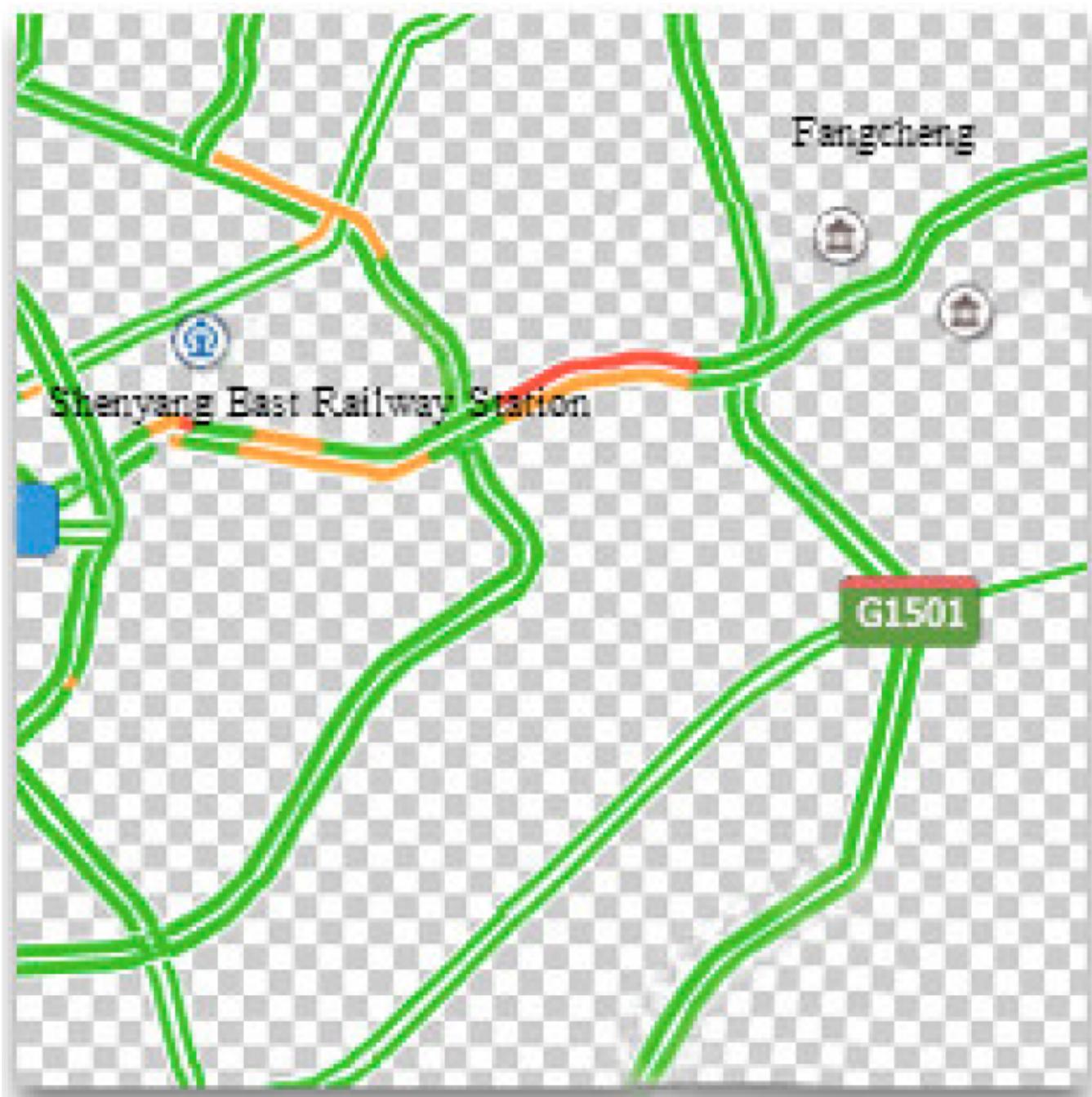


Figure 3: Traffic Congestion[6]

temperature Numeric	humidity Numeric	pressure Numeric	wind Numeric	visibility Numeric	road_length Numeric	tfs Numeric	poi_number Numeric	aqi Nominal
5.5	89.0	758.1	2.0	14.0	2185.0	2371.0	63.0	excellent

Figure 4: Structure of RAQ data[6]

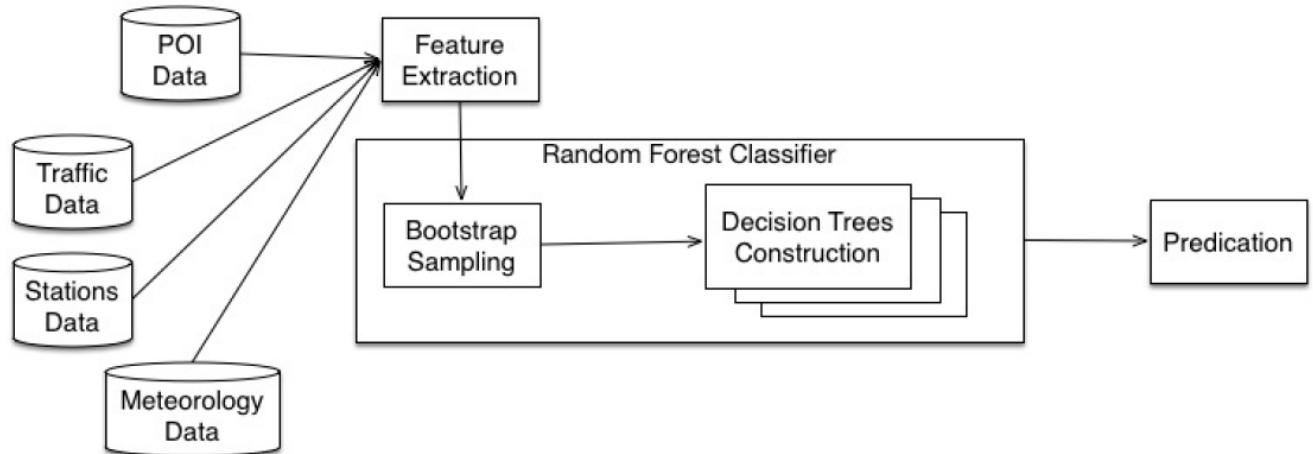


Figure 5: Procedure of RAQ [6]

### Algorithm 1. RAQ

**Input:** A dataset  $S$  with features:  $F_{mt}, F_{mh}, F_{mp}, F_{mw}, F_{mv}, F_{ri}, F_{tcs}, F_{pn}$  and labeled AQI level; unlabeled dataset  $U$ ; trees quantity  $T$ ; features quantity  $m$ ;

**Output:** AQI level

- 1 for  $T$  trees
- 2 randomly select  $m$  features from  $S$ ;
- 3 for  $m$  features in each node
- 4 calculate information gain by Equation (3);
- 5 choose maximum gain to split the dataset in the node;
- 6 remove used feature from feature candidates;
- 7 input unlabeled data into trees;
- 5 get predicted AQI level according to Equations (5) and (6);

Figure 6: RAQ Algorithm [6]

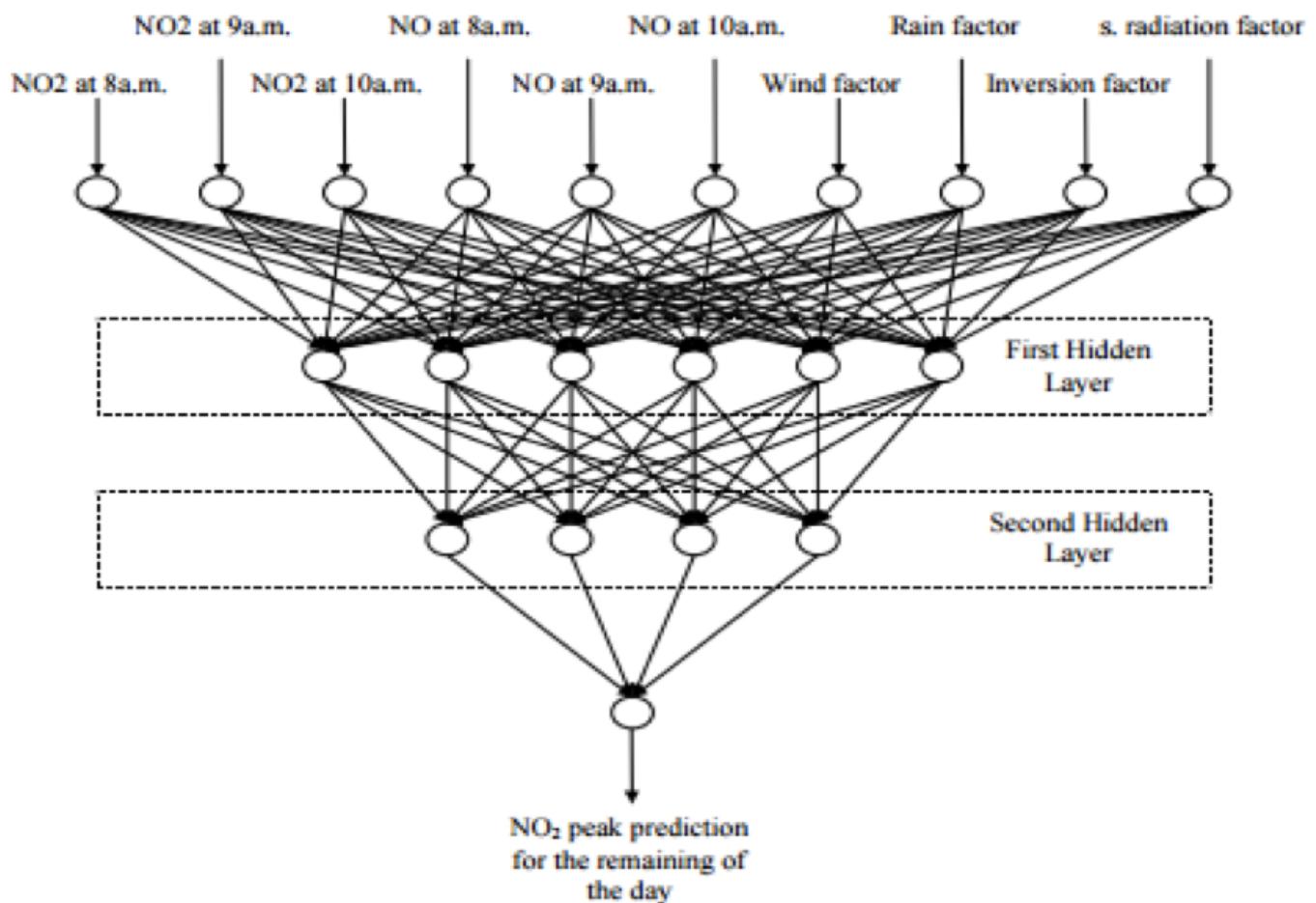


Figure 7: Artificial Neural Network(ANN) Model [4]

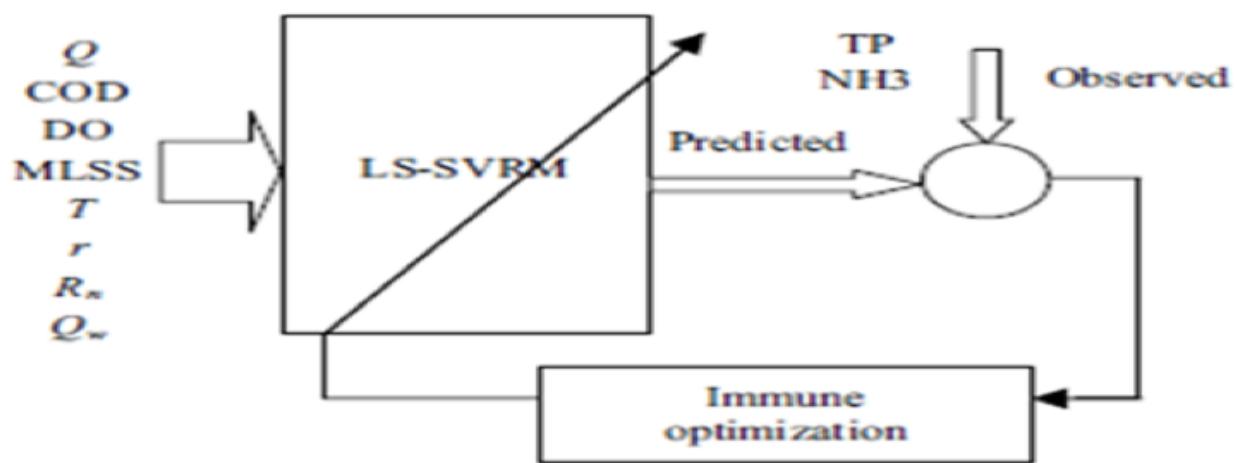


Figure 8: Least squares Support Vector Machine Model [4]

```
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
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-11-06 17.36.21] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.8s.
```

---

```
Compliance Report
```

---

```
name: Vegi, Karthik
hid: 231
paper1: Oct 29 17 100%
paper2: 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
8  
wc 231 paper2 8 565 content.tex  
wc 231 paper2 8 2082 report.pdf  
wc 231 paper2 8 257 report.bib
```

```
find "
```

---

```
102: Do not use "these quotes" but use these ``these quotes''.  
passed: False
```

```
find footnote
```

---

```
112: \footnote{do not use footnotes}.
```

```
passed: False
```

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

---

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

```
passed: True
```

```
floats
```

---

```
62: In Figure \ref{f:fly} we show a fly. Please note that because we  
use
```

```
68: \begin{figure}[!ht]
```

```
69: \centering\includegraphics[width=\columnwidth]{images/fly.pdf}
```

```
70: \caption{Example caption}\label{f:fly}
```

```
85: or generate them by hand while using the provided template in  
Table\ref{t:mytable}. Not ethat
```

```
88: \begin{table}[htb]
```

```
91: \label{t:mytable}
```

```
figures 1  
tables 1  
includegraphics 1  
labels 2  
refs 2  
floats 2
```

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

Label/ref check

```
105: 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 change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
=====
```

---

```
The following tests are optional
```

---

```
=====
```

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

```
find newline
```

---

```
passed: True
```

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

```
find cite {
```

---

```
passed: True
```

# Big Data Analytics in Developing Countries

Judy Phillips  
Indiana University  
PO BOX 4822  
Bloomington, Indiana 47408  
judkphil@iu.edu

## ABSTRACT

Developing nations cope with numerous humanitarian challenges. Infrastructures are often inadequate to deal with basic public health, public safety, and environmental concerns. As a result, citizens deal with issues such as poverty, food insecurity, and the unavailability of basic health care. Resource limitations often make it difficult to manage crisis situations such as natural disasters. The use of wireless and Internet related technology is growing globally. Mobile phone and social media usage are becoming common even in remote areas. As a result, Big Data analytics is playing a role in mitigating the impacts of some of these humanitarian concerns.

## KEYWORDS

I523, HID332, developing countries, food insecurity, public safety, big data

## 1 INTRODUCTION

Individuals in developing nations face a long list of humanitarian challenges, including poverty, hunger, health care access, and availability of clean water sources. Other challenges include insufficient resources to deal with public safety and crisis intervention issues.

The statistics are dismal. "Almost 1.3 billion people living in developing countries live on less than 1.50 dollars a day" [2]. "According to the United Nations, approximately twenty two thousand children die each day in these countries due to poverty" [1]. More than eight hundred seventy million people in third world nations have no food to eat or a very precarious food supply. "A third of all childhood deaths in sub-Saharan Africa are caused by hunger related diseases" [1]. That is approximately 2.6 million deaths per year. One child dies every five seconds of starvation [1]. More than two hundred million children under five years of age in developing countries do not reach their developmental potential due to malnutrition [7]. Over 1.2 billion people around the globe do not have regular access to clean drinking water. Many people die from common curable diseases that such as malaria, pneumonia, and diarrhea because they do not have access to health care. Approximately ten million children die each year from treatable diseases. [2]. Fifty percent of pregnant women in developing countries lack proper prenatal care. This results in over three hundred thousand maternal deaths annually from childbirth. [1]. The threat of HIV is also reaching a pandemic level in many of the third world countries [1].

Big Data Analytics is starting to be used to address some of these issues. Digital data is becoming more widely available globally. Internet wireless communications and mobile phone access are starting to become commonplace even in some rural areas. The

data collected from these devices is being combined with data collected via traditional data sources such as datasets and surveys. This is providing information and insights that have never before been available. "The diffusion of data science into the realm of international development constitutes an opportunity to bring powerful new tools in the fight against poverty, hunger, and disease" [5]. Furthermore, the real time availability of much this data enables more timely and agile implementations of solutions. This all results in significantly better outcomes.

## 2 INFRASTRUCTURE

In recent years, there has been a huge increase in the availability of digital technology globally, including in developing nations. According to the GSM Association 79 percent of the worlds total inhabited areas had mobile network coverage in 2012 [4]. According to the International Telecommunications Union, there were 2 billion people using the internet in 2015 and there were 91.8 mobile phone subscriptions per 100 inhabitants in developing countries [4]. Social media such as Facebook and twitter is being utilized by more and more people worldwide. Sensor technology is becoming less expensive and more efficient. Better algorithms are being developed to utilize the lower cost sensors for developmental activities. Data information sources include call logs, mobile banking transactions, blog posts, tweets, and Facebook content [5].

The diffusion of mobile phone technology has been especially important. Because mobile phones are often the only interactive technology that low income individuals have access to, they have become the cornerstone of many Big Data projects in the developing world[4].

## 3 BIG DATA

The amount of data that is being generated in developed countries is increasing rapidly. According to the Cisco Global Cloud index the highest workload growth rates between 2013 and 2018 are expected to be in the Asian Pacific, the Middle East and Africa, and Latin America. Growth rates during these time periods are expected to be 45 percent, 39 percent, and 34 percent respectively. "Data center traffic in the Middle East and Africa is expected to reach 366 exabytes in 2018 compared to 68 exabytes in 2013" [4].

## 4 HEALTHCARE

"Big Data has enormous potential to address health care challenges in the developing world" [4]. One of the primary problems with healthcare in the developing world is the overall lack of access. This is caused by a combination of geographical accessibility and the lack of basic medical resources. There are shortages trained medical professionals, medical equipment, and drug stocks. People

in rural areas often have to travel long distances in order to obtain care. There are also a lack of resources to implement basic public health regimes such as immunization policies. All of this makes the occurrence of serious disease outbreaks and epidemics common and difficult to manage when they do occur. Another issue is the existence of widespread fake drug distribution networks.

#### 4.1 Public Health

One area in which Big Data can have an enormous impact on the health of vulnerable populations is in public health policy. Proper public health infrastructure is needed to prevent, treat, and manage serious disease outbreaks. Public health policies and related public education can also educate populations and influence attitudes and behaviors concerning important health related matters such as maternal health and immunizations.

Big Data is extremely useful for managing serious disease outbreaks, including pandemics and epidemics. Big Data and data science can be used first to track and monitor the spread of the disease and then to effectively allocate resources and medication so that the disease can be properly treated and contained. In fact, the term for this field is Infodemiology. It is a whole new field of data science [5].

Health related data is mined from social media and sites such as twitter and then combined with data visualization techniques to track the geographic spread of a disease. As the spread of the disease is being tracked in real time, big data is used to ensure that all available resources are allocated effectively. Big data ensures the right distribution of resources, including medical personnel and medication at the right time to the right location [6]. Proper resource allocation is especially important when lifesaving medical supplies are in short supply. According to the US Center for Disease Control and prevention (CDC), online data can help detect disease outbreaks before confirmed diagnosis or lab confirmation [5]. It is estimated that disease outbreaks can be identified up to two weeks sooner than with the use of traditional methods such as physician reporting [4]. When this resource allocation technique was used in Tanzania during a malaria outbreak, it reduced the number of drug facilities that were out of stock of the appropriate medication during the epidemic from 78 percent to 26 percent [4].

Social Media can also be used to track peoples health related beliefs, perceptions and concerns at any a given time and in real time. This methodology is referred to as sentiment analysis. For example, researchers can get an indication of health related attitudes about immunizations, the use of medication or prenatal care programs by reviewing social media posts. These studies can assist with health related education efforts. Social media and big data analytics are also be used to measure the impacts of humanitarian aid and intervention. For example, the United Nations used this technique to evaluate whether the Every Woman Every Child initiative had had an impact. This was a program that was designed to increase awareness of maternal health, breastfeeding, vaccinations. A team of researchers analyzed social media posts for two years for relevant keywords, such as breastfeeding or vaccination to determine if the program has resulted in increased parental awareness [4]. The information collected can be used to identify needs in order to establish and manage public health policies and programs.

Sentiment analysis can also be used to track other public health related issues such housing shortages, employment, and inflated food prices. This methodology is able to identify issues earlier than traditional methods and thus enables more timely deployment of resources and solutions [5].

#### 4.2 Health Care Access

In developing countries, there are often problems with geographical accessibility to health care. People in rural areas often need to travel long distances to visit a health care professional. Also, rural areas do not have enough primary health care providers and specialists are rarely available.

The Internet of Things technology can solve some of these issues. One solution is patient sensors. Relatively low cost sensors can be worn on the person to monitor physiological variables in real time. The data collected can be transmitted to health care providers in a distant locations for diagnosis and treatment. These sensors can be used for routine as well as critical health issues such as heart palpitations. For example, in Africa there is a device called Cardio pad. It is a medical tablet that can be used to perform and collect information from cardiology related tests by individuals who have no cardiac training. The information gathered can then be sent to a cardiac specialist via mobile phone in order to receive diagnosis and treatment instructions[4]. In China the Internet of Things technology Institute is developing a telephone booth sized health capsule. Rural villagers can be receive a diagnosis from a distantly located physician when they step into it. [4].

#### 4.3 Distribution of Fake Drugs

The widespread distribution of fake drugs is a huge health hazard in developing nations. According to the World Health Organization, counterfeit antimalarial and tuberculosis drugs account for seven hundred thousand deaths annually. Big Data technology is playing a huge role in fighting this crime. One nonprofit organization has developed a possible solution. The name of the program is called GoldKeys. All legitimate prescription containers have a twelve digit scratch off code. Customers can verify the authenticity of the medication by texting the scratched off code number to a health hotline. The number is matched to information in a cloud database and the information is sent back to the customer. The project is being maintained and funded primarily by Hewlett Packard [4].

### 5 ENVIRONMENTAL PROTECTION AND WATER SUPPLY

Almost a billion people in the world to not have a reliable source of clean drinking water [1]. According to World Water Development Report in 2012, inadequate sanitation and poor hygiene result in 3.5 million deaths annually [4]. Much of the water is wasted or leaked due to faulty pipes. Other water is lost due to unidentified or unnecessary pollutants.

The Internet of Things can be used for the purpose of monitoring water supply and quality. Sensors are frequently used to monitor pollutants in a river or water source. Resources are deployed to remedy problems when they are detected. One example is in the city of Da Nang, Vietnam. Da Nang is a major port city on the South China Sea. The Da Nang water company uses Big Data to provide

real time analysis of the city's water supply. The goal is to better manage leaks, monitor pollutants, and accurately forecast future demand. Big Data sensors are installed throughout each stage of the water treatment process. Water quality is tracked in real time. Notifications are sent if there are problems. [4].

In another example, IBM worked with the city of Tshwane in South Africa to develop a crowd source application that users use to report water supply issues such as faulty pipes. The result was the discovery of thirty million dollars of wasted water sources. This application operates without the need of a central inspection authority [3].

## 6 PUBLIC SAFETY AND CRISIS INTERVENTION

One of the most important areas in which Big Data is being deployed is to enhance public safety and crisis intervention efforts during natural disasters. "The availability of digital data collected and analyzed rapidly and in real time can drastically improve interventions and outcomes in crisis situations for vulnerable populations" [5].

One of the most widely used tools in this effort are crisis maps. Crisis maps use data from numerous sources, including local citizen reports, social network data, and environmental data to aid emergency responders in times of natural disaster. "Crisis maps have been deployed during dozens of events worldwide, including the 2012 Haiti earthquake and the 2010 Pakistan floods" [3]. In Haiti during an earthquake a centralized text message center was set up that allowed cell phone users to report where people were trapped. The United States Geological Survey has developed a system that monitors Twitter for spikes about earthquakes globally. This information can be used to evaluate the location, quantify magnitude, identify epicenter, and respond quickly and appropriately [5].

## 7 AGRICULTURE

"More than half the population in all of the developing nations depend upon agriculture and farming for at least two meals a day. This accounts for almost seventy five percent of the world's poorest people" [1]. Therefore, one important way to address poverty and food insecurity is to find ways to make farming techniques more effective and productive. Big Data has big potential to dramatically increase production for small scale farmers.

"Studies suggest that ineffective farm operations such as late planting, lack of proper land preparation, improper harvesting techniques and poor housing and feeding of livestock can reduce a smallholders' farmers' productivity by up to forty percent" [4]. One technique for improving production is Precision agriculture. The objective of Precision agriculture is to provide farmers with informed, personalized information so that they can make better operational decisions in real time. Data is collected on things such as soil conditions, weather, seeding rates, and crop yields using technology such as sensors, drones and satellites [4]. Sensors can be located in fields, inside livestock, or on farm equipment. After the data is collected it is analyzed and returned to the farmers via computers and mobile phones in terms of customized solutions. Instructions may be such things as the optimum type of seeds, pesticides, herbicides, and fertilizer use. The objective is to match inputs with the exact need. When resources are used efficiently

production is maximized. Another solution involves collecting data to locate and notify farmers of the spread of crop and livestock plagues. The objective is that farmers take safety measures as soon as possible [3].

In Uganda there is a Big Data tools project that uses Precision agriculture techniques that were developed by the Grameen Foundation. Data is collected on farmers, farming practices, and external conditions. It is given back to farmers in the form of a community knowledge database via Android phones. Information about the time and methods of planting crops, caring for farm animals and marketing their products [4].

Another way in which big data can be used for small holder farmers to support financing opportunities. In Nairobi, Africa the company Gro Ventures is building a platform which integrates information about crops and the environmental conditions to give lenders more confidence to lend money to farmers. One of the offerings allows farmers to pool their data to apply for collective loans to buy shared tractors and equipment [3].

## 8 CHALLENGES

There are many challenges to the successful implementation of many of these projects. Many people in the least developed nations still lack access to internet service or a mobile phone. There are high costs associated with using big data technology. Cost of mobile phones, analytical services and data services often cost prohibitive for individual citizens. There is also a Big Data skill set deficient. Big data technology and the analytics to turn big data into actionable information requires technical skills that are often not available. Furthermore, health care professionals and other related personnel often lack knowledge or training about data science.

In order for initiatives to be successful, financial and technical support will need to come from other sources: academia, public and private sector, and philanthropic. To date, there are numerous non-government organizations (NGOs) working throughout the world to fight poverty and reduce disease [6]. The United Nations started an initiative in 2009 called Global Pulse. The objective of Global Pulse is to research ways that Big Data can be incorporated into the developing world to improve lives. They are currently conducting several research initiatives in various locations throughout the world. Several private organizations are also playing a role. For example, Google has announced a plan to develop high speed internet solutions in developing countries using high altitude balloons. Their goal is to add an additional 1 billion people to the Internet from Africa, and Southwest Asia [4].

## 9 CONCLUSION

Although Big Data does not have the ability to solve all of the world's problems, it does have enormous potential to reduce suffering and save lives for those living in developing countries. Big data is giving smallholder farmers resources to substantially increase their food production. This will play a substantial role in the fight against poverty and food insecurity. Big data analytics is improving health by making health care accessible to even those in the most remote locations. Big data provides the knowledge to identify and monitor water availability issues such as waste and pollution so that problems can be identified and dealt with immediately. Big is

also saving lives by providing the real time knowledge needed to respond effectively to health epidemics and natural disasters. As the use of internet related devices continues to increase throughout the developing world, the impact of big data will continue to grow.

## 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] ELIST10. 2014. Top 10 Major Problems in Third World Countries. Web page as Article. (June 2014). <http://www.elist10.com/top-10-major-problems-third-world-countries/>
- [2] Institute of Ecolonomics. 2015. Top 5 Challenges the Third World is Facing Today. Web page as Blog. (May 2015). <http://ecolonomics.org/top-5-challenges-the-third-world-is-facing-today/>
- [3] Travis Korte. 2014. How Data Analytics Can Help the Developing World. Web page as Article. (Sept. 2014). [https://www.huffingtonpost.com/travis-korte/how-data-and-analytics-ca\\_b\\_5609411.html](https://www.huffingtonpost.com/travis-korte/how-data-and-analytics-ca_b_5609411.html)
- [4] Nir Kshetri. 2016. *Big Data's Big Potential in Developing Economies*. CABI, Wallingford Oxfordshire, UK.
- [5] United Nations Global Pulse. 2012. Big Data for Development Challenges and Opportunities. Web page as paper. (May 2012). <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseMay2012.pdf>
- [6] Mark Van Rijmenam. 2017. How Big Data Analytics Can Help the Developing World Beat Poverty. Web page as Article. (July 2017). <https://datafloq.com/read/big-data-developing-world-beat-poverty/168>
- [7] Wikipedia. 2017. Developing Country. Web page. (Oct. 2017). [https://en.wikipedia.org/wiki/Developing\\_country](https://en.wikipedia.org/wiki/Developing_country)

```
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
I was expecting a ',' or a '}'---line 7 of file report.bib.bib
:
: isbn      = {1 3; 978 1 78064 868 2},
(Error may have been on previous line)
I'm skipping whatever remains of this entry
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-11-06 17.38.04] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Typesetting of "report.tex" completed in 1.0s.
```

---

```
Compliance Report
```

---

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

```
yamlcheck
```

---

```
wordcount
```

---

```
4
wc 332 paper2 4 2991 report.tex
wc 332 paper2 4 2983 report.pdf
wc 332 paper2 4 302 report.bib
```

find "

---

- 35: The statistics are dismal. "Almost 1.3 billion people living in developing countries live on less than 1.50 dollars a day" \cite{www-google-top5}. "According to the United Nations, approximately twenty two thousand children die each day in these countries due to poverty" \cite{www-google-top10}. More than eight hundred seventy million people in third world nations have no food to eat or a very precarious food supply. "A third of all childhood deaths in sub-Saharan Africa are caused by hunger related diseases"\cite{www-google-top10}. That is approximately 2.6 million deaths per year. One child dies every five seconds of starvation \cite{www-google-top10}. More than two hundred million children under five years of age in developing countries do not reach their developmental potential due to malnutrition \cite{www-google-WikiDevC}. Over 1.2 billion people around the globe do not have regular access to clean drinking water. Many people die from common curable diseases that such as malaria, pneumonia, and diarrhea because they do not have access to health care.
- Approximately ten million children die each year from treatable diseases. \cite{www-google-top5}. Fifty percent of pregnant women in developing countries lack proper prenatal care. This results in over three hundred thousand maternal deaths annually from childbirth. \cite{www-google-top10}. The threat of HIV is also reaching a pandemic level in many of the third world countries \cite{www-google-top10}.
- 37: Big Data Analytics is starting to be used to address some of these issues. Digital data is becoming more widely available globally. Internet wireless communications and mobile phone access are starting to become commonplace even in some rural areas. The data collected from these devices is being combined with data collected via traditional data sources such as datasets and surveys. This is providing information and insights that have never before been available. "The diffusion of data science into the realm of international development constitutes an opportunity to bring powerful new tools in the fight against poverty, hunger, and disease" \cite{www-google-GloPls}. Furthermore, the real time availability of much this data enables more timely and agile implementations of solutions. This all results in significantly better outcomes.
- 45: The amount of data that is being generated in developed countries is increasing rapidly. According to the Cisco Global Cloud index the highest workload growth rates between 2013 and 2018 are

expected to be in the Asian Pacific, the Middle East and Africa, and Latin America. Growth rates during these time periods are expected to be 45 percent, 39 percent, and 34 percent respectively. "Data center traffic in the Middle East and Africa is expected to reach 366 exabytes in 2018 compared to 68 exabytes in 2013" \cite{DevEcon}.

- 49: "Big Data has enormous potential to address health care challenges in the developing world" \cite{DevEcon}. One of the primary problems with healthcare in the developing world is the overall lack of access. This is caused by a combination of geographical accessibility and the lack of basic medical resources. There are shortages trained medical professionals, medical equipment, and drug stocks. People in rural areas often have to travel long distances in order to obtain care. There are also a lack of resources to implement basic public health regimes such as immunization policies. All of this makes the occurrence of serious disease outbreaks and epidemics common and difficult to manage when they do occur. Another issue is the existence of widespread fake drug distribution networks.
- 80: One of the most important areas in which Big Data is being deployed is to enhance public safety and crisis intervention efforts during natural disasters. "The availability of digital data collected and analyzed rapidly and in real time can drastically improve interventions and outcomes in crisis situations for vulnerable populations" \cite{www-google-GloPls}.
- 82: One of the most widely used tools in this effort are crisis maps. Crisis maps use data from numerous sources, including local citizen reports, social network data, and environmental data to aid emergency responders in times of natural disaster. "Crisis maps have been deployed during dozens of events worldwide, including the 2012 Haiti earthquake and the 2010 Pakistan floods" \cite{www-google-Hffpst}.
- 87: "More than half the population in all of the developing nations depend upon agriculture and farming for at least two meals a day. This accounts for almost seventy five percent of the worlds poorest people" \cite{www-google-top10}. Therefore, one important way to address poverty and food insecurity is to find ways to make farming techniques more effective and productive. Big Data has big potential to dramatically increase production for small scale farmers.
- 89: "Studies suggest that ineffective farm operations such a late

planting, lack of proper land preparation, improper harvesting techniques and poor housing and feeding of livestock can reduce a smallholders farmers productivity by up to forty percent"  
\cite{DevEcon}.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

floats

---

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

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

Label/ref check

passed: True

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

find textwidth

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib.bib
I was expecting a ',' or a '}'---line 7 of file report.bib.bib
:
:    isbn      = {1 3; 978 1 78064 868 2},
(Error may have been on previous line)
I'm skipping whatever remains of this entry
(There was 1 error message)
```

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

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

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Analysis for Computer Network Defense

Jordan Simmons  
Indiana University  
Smith Research Center  
Bloomington, IN 47408, USA  
jomsimm@iu.edu

## ABSTRACT

Computer security threats and attacks are constantly evolving. Everyday, hackers are creating new techniques to bypass network security for the purpose of malicious attacks. To keep up with the changing intrusion technologies, the technologies that defend these attacks need to constantly evolve also. Modern day technologies use deep learning techniques to monitor network activity, and detect malicious code. We will provide an overview of network security and modern technologies being used to protect computer systems and networks.

## KEYWORDS

i523,HID336, Computer Network Security, Big Data Analysis, Deep Learning, Intrusion Detection Systems,

## 1 INTRODUCTION

Everyday a different computer network is being breached with the intent to cause harm to the system or to steal valuable data. Computer hackers are constantly creating new ways to evade network security and create malicious code that can not be detected by security systems. As malicious technologies continue to advance, the technologies that defend against these technologies need to adapt with these advances. The problem with computer network defence is that the technologies used to breach systems constantly change. Once a solution is created to defend a technology, a new malicious technology could be created the next day. Today many security specialist are using deep learning technologies to monitor network intrusions, and detect malicious code. In order to better understand computer network defense, an overview of modern attacks, network data collection processes, and the technologies used to analyze network data is provided.

## 2 DATA COLLECTION

### 2.1 Network Intrusion Data Collection

### 2.2 Malware Data Collection

## 3 DEEP LEARNING FOR NETWORK INTRUSIONS

## 4 DEEP LEARNING ON MALWARE

## 5 CONCLUSION

## 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 do this change TODO with DONE. However if you check something on with DONE, but we find you actually have not executed it correctly, 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.

### 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 inflate 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 float. 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

```
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
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-11-06 17.38.24] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Empty 'thebibliography' environment.
Missing character: ""
Typesetting of "report.tex" completed in 1.6s.
./README.yml
 8:81      error    line too long (84 > 80 characters) (line-length)
 9:81      error    line too long (85 > 80 characters) (line-length)
10:81      error    line too long (82 > 80 characters) (line-length)
11:81      error    line too long (81 > 80 characters) (line-length)
12:52      error    trailing spaces (trailing-spaces)
26:81      error    line too long (82 > 80 characters) (line-length)
26:82      error    trailing spaces (trailing-spaces)
29:79      error    trailing spaces (trailing-spaces)
31:62      error    trailing spaces (trailing-spaces)
33:79      error    trailing spaces (trailing-spaces)
```

---

```
Compliance Report
```

---

```
name: Jordan Simmons
hid: 336
paper1: Oct 25 17
```

```
paper2: In Progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
2
```

```
wc 336 paper2 2 457 report.tex  
wc 336 paper2 2 1097 report.pdf  
wc 336 paper2 2 50 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

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

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

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

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

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I found no \citation commands---while reading file report.aux
Database file #1: report.bib
(There was 1 error message)
```

```
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
```

# Big Data Analytics and IoT Smart Refrigerators

Robert W. Gasiewicz

Indiana University

711 N. Park Avenue

Bloomington, IN 47408

rgasiewi@iu.edu

## ABSTRACT

The intent of this paper is to explore the rapid growth of IoT Smart Appliances, specifically with regard to refrigerators. As more devices are connected to the internet, to each other, and become readily available to consumers, there are many exciting new possibilities that offer both convenience and to make our lives more efficient. The scope of this paper will begin with a brief history of IoT, then move on to describe the current way in which this technology is being applied, and conclude with exploration and outlook on future development possibilities as well as potential risks.

## KEYWORDS

i523, HID316, Big Data, IoT, Refrigerators, Smart Appliances, M2M, Samsung, Innit, Instacart, GrubHub

## 1 INTRODUCTION

The advent of the Internet of Things (IoT) began at the close of the last millennium when the world began connecting ordinary devices - electronics other than traditional computers - to the internet. With virtually unlimited possibilities, the unthinkable became reality when the concept of putting a wireless network card in a refrigerator went mainstream. Initial features were as simple as a large touchscreen with the news, the weather, and a doodle board.

From there, IoT Smart Refrigerators have evolved to become equipped with cameras, cooking recommendations, and even rudimentary food inventory and spoilage management systems. Now that food delivery services such as Instacart and GrubHub have become popular, there are already plans to integrate these services with smart refrigerators. As IoT has continued to expand throughout the marketplace and the concept of machine-to-machine (M2M) IoT has taken hold, there are now even more possibilities, which means a bright future in the kitchen no matter if you're an aspiring chef, a person trying to efficiently manage a family, or someone with specific health needs. However, along with the rapid advance of new features, there are also significant threats and blind spots with security.

## 2 EARLY HISTORY OF IOT AND NETWORKED APPLIANCES

Although the internet didn't yet exist in the minds of Hollywood producers in 1985, the opening scene from Back to the Future begins with a room full of ticking clocks, one of which is an alarm clock that rings and sets off a Rube Goldberg machine that has been configured by Doc Brown to automate the preparation of his breakfast. It's not unreasonable to believe that, in his many time

travel escapades, Doc would've eventually *discovered* the internet and would've upgraded this rudimentary appliance.

That reality wouldn't come until five years later in 1990 when the first IoT device, a toaster, was turned off and on via the internet. At the October 1989 INTEROP Conference, John Ramkey used a Sunbeam Radiant Control toaster connected to a TCP/IP network to demonstrate that the device could be turned off and on [11]. Not only did Ramkey succeed at turning the toaster on and off, he used SNMP code delivered via his computer's parallel port to a larger relay to control power to the toaster. The SNMP code executed commands for a value, 1 through 10, for the toast's doneness as well as a calculation for the type of item being toasted. For example, while the command for wheat bread would tell the toaster to toast at a level of 2, the command for a frozen bagel would tell the toaster to toast at a level of 5. Additional innovations were later added, such as a Lego robotic arm to insert the bread into the toaster; a sight Doc Brown would've been proud to see.

By 1999, the Salt Lake City Tribune/Deseret News was predicting that household appliances like the refrigerator were going to be part of a future in which "everyone lived like the Jetsons" [12]. "The networked home is on the horizon", the Tribune/Deseret News' Michael Stroh wrote, "with a click, you call up your refrigerator on your office PC to see what's inside (a bar-code reader within the fridge keeps a running inventory). The refrigerator suggests lasagna but warns that you'll need to buy ricotta - and a few other items" [12]. Not surprisingly, it would be at least another decade before this concept became a viable reality.

## 3 IOT IS BORN

The first time the term *Internet of Things* was used wasn't until nine years later by Kevin Ashton, co-founder of the Auto-ID Center at the Massachusetts Institute of Technology (MIT). The Auto-ID Center was founded with the expressed purpose of creating a formal standard for Radio Frequency Identification (RFID) and other types of networked sensors. In 2009, Kevin wrote [5]: "I could be wrong, but I'm fairly sure the phrase *Internet of Things* started life as the title of a presentation I made at Procter & Gamble (P&G) in 1999. Linking the new idea of RFID in P&G's supply chain to the then-red-hot topic of the Internet was more than just a good way to get executive attention. It summed up an important insight which is still often misunderstood."

Even though Kevin briefly captured the momentary attention of the C-Suite at P&G, it wasn't another full decade until the true concept of IoT caught on in the marketplace. In 2011, the market research company Gartner, included IoT on their hype cycle chart for the very first time. By 2016, IoT was past-peak of inflated expectations was doing the usual nosedive into the trough of disillusionment [7].

[Figure 1 about here.]

## 4 IOT SMART REFRIGERATORS COME OF AGE

Internet refrigerators, on the other hand were a bit slower catching up. After many failed attempts in the mid-2000s at various gimmicky models, it seemed that the once rosy future painted by Mr. Stroh a decade earlier was simply not going to come to fruition. Hardware and network technology had not yet caught up. By 2014, murmurs of a new wave of internet fridges hit the marketplace and excitement began to build, and by 2016, the IoT Refrigerator was ready for primetime. On January 24, 2016, Samsung launched its Smart Hub Refrigerator complete with a massive 21.5 inch 1080P touchscreen and Android operating system. Another exciting new feature of the Smart Hub fridge was the interior cameras that allowed users to get a real-time look at the contents of their fridge from anywhere[9].

A year later, Samsung debuted version 2.0 of the Smart Hub fridge, this time with improvements such as third-party apps such as Spotify and individualized user profiles for family members. Users are also able to serve photos and other content to the screen as well. Interestingly, Samsung has opted to go with its own proprietary voice control system called S-Voice, while its only current competitor in the IoT fridge marketplace, LG, will integrate with Amazon's Alexa. Only in Europe, with the Lidl supermarket chain, will consumers be able to order groceries through the the fridge. It's a start, but there is much, much more on the horizon[8].

## 5 THE FUTURE OF IOT SMART REFRIGERATORS

The future of IoT Smart Refrigerators - and kitchen appliances working in concert in general - is brighter than perhaps Doc Brown or even John Ramkey could have ever imagined. Hardware, networking, and most importantly, software, have all caught up to be viable in fulfilling consumer demands and there are fresh new ideas already just beginning to hit the marketplace. The next phase of the IoT Smart Refrigerator will be one that is marked by progress in software. Structurally speaking, refrigerators are designed to last between 14-17 years[2], however, the average consumer might upgrade their personal computer 3 to 4 times during this time span. In other words, an IoT Smart Refrigerator made today, might only be 1/4 to 1/3 of the way through its average lifespan before its computer and networking components become obsolete.

One Silicon Valley company that seems to have a viable solution to this problem is Innit[4]. Innit has come up with the idea of having a cloud-based platform for the kitchen that partners with appliance manufacturers such as Jenn Air and Whirlpool to add their components and integrate their application with existing appliance platforms. The idea is that you can equip your entire kitchen, not just the refrigerator, with technology that can make anyone a culinary master with a bit of guidance[10]. Building upon Samsung's successful Smart Hub fridge platform, Innit takes the camera-in-your-fridge concept a step further by introducing image recognition software that can be used to interface with the cloud to generate recipes based on available ingredients, manage spoilage,

and inventory - including placing orders for new food. The technology would also enable other kitchen appliances such as an oven or microwave to interact with one another to create a meal.

Aside from personal convenience, one of the most significant values derived from the advance of this sort of technology is that it could prevent an enormous amount of food waste. The United Nations' Food and Agriculture Organization estimates that up to 1.3 billion tons of food are wasted globally every year[3], which equates to roughly 30 percent of all food produced in the same time-frame. Ultimately, software like Innit's because it is connected to the cloud and utilizing big data to allow consumers to make informed decisions about what they eat, people will live and eat healthier and greener.

## 6 SMART AND DANGEROUS: AN IOT DOUBLE-EDGED SWORD

Yes - it is true - both today and in the future, your IoT Smart Refrigerator will help you live better, but as Swapnil Bhartiya points out in a recent article on InfoWorld[6], it could also kill you. It sounds ominous, but the rapid growth of IoT comes with a steep price: lack of security. Consumers can never really be sure if their software will be patched properly and for how long. It has been well-documented that hackers have been able to successfully commandeer smart devices and utilize them to aggressively launch DDoS that disabled a sizeable portion of the internet. An even bigger threat is that, once compromised, a vulnerable smart device will work as a Trojan Horse allowing nefarious users to access other devices on your local network. Once you throw Alexa into the mix, all bets are off.

One development that is offsetting this risk is the unification of IoT networks in the cloud. Samsung is now creating a SmartThings cloud in which all of its IoT devices will interact. This centralization makes security and big data much easier to manage. This unification is also occurring at the macro level with Cisco and Google's cloud[1] which will hopes to achieve the following goals:

- (1) Freedom to access any resource while preserving security and compliance
- (2) Ability to extend policy to cloud environments to optimize applications
- (3) Extend visibility, threat detection and control across hybrid environments without slowing innovation

## 7 CONCLUSION

IoT has a very bright future ahead and the rapidly evolving IoT Smart Refrigerator will serve as the centerpiece not only to a smart, connected kitchen, but to a smart, connected, and secure home. While it was hardware and networking that delayed progress in the 1990s and software and implementation that led to stagnation in the 2000s, security serves as the next challenge to be overcome as IoT Smart Refrigerators join the burgeoning global network of IoT smart devices.

## REFERENCES

- [1] 2017. Cisco and Google Cloud. (2017). Retrieved October 30th, 2017 from <https://www.cisco.com/c/en/us/solutions/strategic-partners/google-cloud.html>
- [2] 2017. The Expected Life of a Refrigerator. (2017). Retrieved October 30th, 2017 from <http://homeguides.sfgate.com/expected-life-refrigerator-88577.html>

- [3] 2017. Food and Agriculture Organization of the United Nations: Food Loss and Food Waste. (2017). Retrieved October 30th, 2017 from <http://www.fao.org/food-loss-and-food-waste/en/>
- [4] 2017. Innit. (2017). Retrieved October 30th, 2017 from <http://www.innit.com>
- [5] Kevin Ashton. 2009. That 'Internet of Things' Thing. *RFID Journal* (jun 2009), 1. <http://www.rfidjournal.com/articles/view?4986>
- [6] Swapnil Bhartiya. 2017. Your smart fridge may kill you: The dark side of IoT. (2017). Retrieved October 30th, 2017 from <https://www.infoworld.com/article/3176673/internet-of-things/your-smart-fridge-may-kill-you-the-dark-side-of-iot.html>
- [7] Inc. Gartner. 2017. Technologies Underpin the Hype Cycle for the Internet of Things, 2016. (2017). Retrieved October 30th, 2017 from <https://www.gartner.com/smarterwithgartner/7-technologies-underpin-the-hype-cycle-for-the-internet-of-things-2016/>
- [8] Rik Henderson. 2017. Samsung Family Hub 2.0 refrigerator preview: Spotify and sausages. (2017). Retrieved October 30th, 2017 from <http://www.pocket-lint.com/review/139892-samsung-family-hub-2-0-refrigerator-preview-spotify-and-sausages>
- [9] Stuart Miles. 2016. Samsung Family Hub Refrigerator comes with giant 21.5-inch screen and camera to spy on your food. (2016). Retrieved October 30th, 2017 from <http://www.pocket-lint.com/news/136305-samsung-family-hub-refrigerator-comes-with-giant-21-5-inch-screen-and-camera-to-spy-on-your-food>
- [10] Rohini Nambiar. 2016. Smart kitchens are a new phase in the Internet of Things, as Innit explains. (2016). Retrieved October 30th, 2017 from <https://www.cnbc.com/2016/07/26/smart-kitchens-are-a-new-phase-in-the-internet-of-things-as-innit-explains.html>
- [11] John Ramkey. 2016. Toast of the IoT: The 1990 Interop Internet Toaster. *IEEE* 6, Article 1 (dec 2016), 3 pages. <https://doi.org/10.1109/MCE.2016.2614740>
- [12] Michael Stroh. 1999. Network systems allow us to live more like the Jetsons. (1999). Retrieved October 30th, 2017 from <https://news.google.com/newspapers?nid=336&dat=19990116&id=lu8jAAAAIBAJ&sjid=iewDAAAIBAJ&pg=3607,488766&hl=en>

LIST OF FIGURES

1 2016 Gartner Hype-Cycle Chart.

5

image missing

**Figure 1: 2016 Gartner Hype-Cycle Chart.**

bibtext report

=====

This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--no key, author in SFGate2017  
Warning--no author, editor, organization, or key in SFGate2017  
Warning--to sort, need author or key in SFGate2017  
Warning--no key, author in Innit2017  
Warning--no author, editor, organization, or key in Innit2017  
Warning--to sort, need author or key in Innit2017  
Warning--no key, author in FAO2017  
Warning--no author, editor, organization, or key in FAO2017  
Warning--to sort, need author or key in FAO2017  
Warning--no key, author in Cisco2017  
Warning--no author, editor, organization, or key in Cisco2017  
Warning--to sort, need author or key in Cisco2017  
Warning--no key, author in Cisco2017  
Warning--no key, author in Cisco2017  
Warning--no key, author in FAO2017  
Warning--no key, author in FAO2017  
Warning--no key, author in Innit2017  
Warning--no key, author in Innit2017  
Warning--no key, author in SFGate2017  
Warning--no key, author in SFGate2017  
Warning--no key, author in Cisco2017  
Warning--no author, editor, organization, or key in Cisco2017  
Warning--empty author in Cisco2017  
Warning--no key, author in SFGate2017  
Warning--no author, editor, organization, or key in SFGate2017  
Warning--empty author in SFGate2017  
Warning--no key, author in FAO2017  
Warning--no author, editor, organization, or key in FAO2017  
Warning--empty author in FAO2017  
Warning--no key, author in Innit2017  
Warning--no author, editor, organization, or key in Innit2017  
Warning--empty author in Innit2017  
Warning--no number and no volume in Ashton01  
Warning--numpages field, but no articleno or eid field, in Ashton01  
(There were 34 warnings)

bibtext \_ label error

=====

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-11-06 17.37.30] pdflatex report.tex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
```

```
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.2s.
```

---

```
Compliance Report
```

---

```
name: Robert Gasiewicz
hid: 316
paper1: 100% Oct 25 17
paper2: 100%
project: 10%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
4
wc 316 paper2 4 1898 report.tex
wc 316 paper2 4 2046 report.pdf
wc 316 paper2 4 408 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
52: \begin{figure}
```

```
54: \%includegraphics[width=\linewidth]{gartner2016.png}
```

```
56: \label{fig:Gartner2016}
```

```
figures 1
```

```
tables 0
```

```
includegraphics 1
```

```
labels 1
```

```
refs 0
```

```
floats 1
```

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

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

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

```
False : check if all figures are referred to: (refs >= labels)
```

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

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

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

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--no key, author in SFGate2017
Warning--no author, editor, organization, or key in SFGate2017
Warning--to sort, need author or key in SFGate2017
Warning--no key, author in Innit2017
Warning--no author, editor, organization, or key in Innit2017
Warning--to sort, need author or key in Innit2017
Warning--no key, author in FA02017
Warning--no author, editor, organization, or key in FA02017
Warning--to sort, need author or key in FA02017
Warning--no key, author in Cisco2017
Warning--no author, editor, organization, or key in Cisco2017
Warning--to sort, need author or key in Cisco2017
Warning--no key, author in Cisco2017
Warning--no key, author in Cisco2017
Warning--no key, author in FA02017
Warning--no key, author in FA02017
Warning--no key, author in Innit2017
Warning--no key, author in Innit2017
Warning--no key, author in SFGate2017
Warning--no key, author in SFGate2017
Warning--no key, author in Cisco2017
Warning--no author, editor, organization, or key in Cisco2017
Warning--empty author in Cisco2017
Warning--no key, author in SFGate2017
Warning--no author, editor, organization, or key in SFGate2017
Warning--empty author in SFGate2017
Warning--no key, author in FA02017
Warning--no author, editor, organization, or key in FA02017
Warning--empty author in FA02017
Warning--no key, author in Innit2017
```

```
Warning--no author, editor, organization, or key in Innit2017
Warning--empty author in Innit2017
Warning--no number and no volume in Ashton01
Warning--numpages field, but no articleno or eid field, in Ashton01
(There were 34 warnings)
```

#### bibtex\_empty\_fields

---

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

#### find ""

---

```
9: number      =      "",  
14: acmid       =      "",  
15: note        =      "",  
22: articleno   =      "",  
24: volume      =      "",  
25: number      =      "",  
28: doi         =      "",  
30: acmid       =      "",  
31: note        =      "",  
39: month       =      "",  
48: month       =      "",  
57: month       =      "",  
66: month       =      "",  
71: author      =      "",  
75: month       =      "",
```

```
80: author =      "",  
84: month =      "",  
93: month =      "",  
98: author =      "",  
102: month =      "",  
111: month =      "",  
116: author =      "",  
120: month =      "",  
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 Using Regression Techniques

Nisha Chandwani

Indiana University Bloomington

Bloomington, Indiana 47405

nchandwa@iu.edu

## ABSTRACT

While analyzing large volumes of data, it is important to make sure that the data is analyzed accurately and efficiently for successful decision making. Predictive modeling has been one of the most critical aspects of analyzing big data. However, the traditional methods of predictive modeling cannot be directly applied to big data as applying statistical analysis to a large volume of data at once is a huge challenge in itself. We discuss how the traditional predictive methods, such as linear regression, can be modified for effectively modeling big data. We then discuss the distributed frameworks like Hadoop and Spark which help in predictive modeling of big data as well as the support extended by programming languages like R for these frameworks. Finally, we provide an overview of some of the regression techniques that can be applied for analyzing big data.

## KEYWORDS

i523, HID203, Big Data, Spark, Hadoop, Predictive Modeling, Regression

## 1 INTRODUCTION

With increasing data from various sources, we have landed in the era of big data. However, collecting big data is just the first step. Effectively using this data for deriving useful business insights is of prime importance. Statistics is the art of learning from data for making optimal business decisions [2]. This learning from data involves the application of statistical methods like regression or time series modeling [2]. The present statistical techniques focus on deriving inference about the population based on sample data. Applying statistical methods to big data in one pass is a huge challenge and thus a more effective method is to partition big data into multiple samples. The results from all the samples can then be used to generate the final result for the predictive model. We present this method in detail in the following section and also discuss some of the regression techniques that can be effectively applied using this divide-and-conquer method on big data.

## 2 TRADITIONAL PREDICTIVE MODELING

Predictive modeling is one of the most important types of statistical analysis of big data. It aims to model the causal relationship between different features present in the data. Specifically, we try to predict the value of one variable, known as the dependent or response variable, with the help of one or more variables, known as independent or explanatory variables. The traditional predictive modeling process is shown in Figure 1.

[Figure 1 about here.]

Each of these steps in traditional predictive modeling is discussed in the following sections [3].

### 2.1 Define Goal

For any predictive modeling, it is important to clearly define the goal, i.e., define the response variable and the explanatory variables that we are going to use to predict the response variable [3].

### 2.2 Data Collection and Management

This is another critical step for predictive modeling which requires identifying the data that can be used for analysis. It can be the most time-consuming step in the entire modeling process and may require some preliminary data exploration and visualization [3]. It also involves identifying the feature set and the structure of each feature.

### 2.3 Data Preprocessing

Before building a predictive model, it is important to check the quality of data. Two major components of data preprocessing are [3]:

- Analyzing missing values: For missing values, it is important to identify whether we want to drop the missing entries in the data or impute them using standard imputation methods such as mean imputation for quantitative features and mode imputation for qualitative features.
- Data transformation: The aim of data transformation is to convert it into a form which is easier to model. For example, normalization and standardization of data help in the better interpretation of the coefficients of the regression model. Some of the transformations also depend on the predictive model that we intend to apply. For example, for a linear regression model, it is important to ensure that the dependent and the independent variables have a linear relationship and that the dependent variable has a constant variance.

### 2.4 Exploratory Data Analysis (EDA)

As part of EDA, we try to summarize the data graphically and analyze each feature along with the relationship between different features. For summarization, a variety of summary statistics such as mean, median, variance, etc. are used [3]. In case of predictive modeling, one might want to explore the data and visualize the numerical summary along with the correlation of different features in the data.

### 2.5 Model-building and reporting

The final step involves building the predictive model using the clean transformed data. Different regression techniques such as linear regression can be used for predicting the response variable from the independent variables. Basically, we try to estimate the coefficient,  $\hat{\beta}$ , for each independent variable such that a unit change

in the independent variable results in a change of  $\hat{\beta}$  units in the mean value of the response variable. Once the model is built, it is evaluated using one of the several metrics like the coefficient of determination.

### 3 PREDICTIVE MODELING WITH BIG DATA

After having the background of the traditional predictive modeling, we now study the process of extending this method for big data. Traditional statistical analysis generally rely on a representative sample of data to make inference about the population [3]. However, in case of big data, relying on one sample may lead to incorrect predictions. Thus, we partition big data into subsets such that the size of these subsets is close enough to a sample. We achieve this by using one of the sampling techniques from statistics such as random sampling, stratified sampling or cluster sampling [2]. We then apply regression techniques to each of these samples independently. The final prediction is made by aggregating the results from each of these samples. The architecture for this divide-and-conquer regression analysis is as shown in Figure 2. Except for the model building step, all other steps are the same as that of traditional predictive modeling.

[Figure 2 about here.]

This divided regression analysis is summarized as below [2]:

- Divide big data: In this step, the data is divided into  $M$  subsets by using one of the statistical sampling techniques.
- Apply multiple linear regression analysis: Perform regression on each of these subsets and compute their respective regression parameters,  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m$ . Combine each of these parameters to compute the final regression coefficient,  $\hat{\beta}_c$  as below:

$$\hat{\beta}_c = f_c(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m)$$

Here,  $f_c$  is a combine function used for combining the coefficients obtained from each of the  $M$  subsets of data. This function can be defined in various ways based on the data. For example, we can define the combine function as the mean value of all the co-efficients:

$$\hat{\beta}_c = \frac{1}{M} \sum_{i=1}^M \hat{\beta}_i$$

- Evaluate the model: As part of model evaluation, we first check the confidence interval for all the co-efficients, obtained by applying the combine function, for all the independent variables. Next, we check the accuracy of the model by using some evaluation metric that measures how accurately the model predicts the value of the response variable. For example, we can use mean squared error ( $MSE$ ) which is calculated as below:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

### 4 BIG DATA SYSTEMS AND THE SUPPORT IN R

As explained in the previous section, regression analysis of big data can be efficiently performed by a divide-and-conquer strategy. However, it is important to remember that big data processing deals with a large volume of data and thus even in case of divided regression analysis, a single machine might not be enough. Thus,

we require a distributed system where subsets of data are processed on multiple machines and the results are later aggregated to produce the final prediction as shown in Figure 3 [4].

[Figure 3 about here.]

Some of the available frameworks that support such distributed storage and parallel computing include Hadoop and Spark. We can use the power of these frameworks to achieve the distributed version of traditional predictive modeling techniques. Many programming languages support these data-intensive computing paradigm. We discuss these frameworks and their support in one of the programming languages, i.e., R [1].

#### 4.1 Hadoop

Hadoop is an open-source Java-based distributed computing platform which is used for distributed storage and distributed processing of huge data sets on computer clusters [1]. All the modules of the Hadoop framework are fault-tolerant, i.e., they can automatically handle failures of individual machines in the framework. The four important modules of this framework are [1]:

- Hadoop Common provides all the Java libraries, utilities, OS level abstraction and the necessary files required by other modules
- Hadoop Distributed File System (HDFS) for storing big data as well as providing high throughput access to this data
- MapReduce for processing large volumes of data
- Hadoop YARN for resource management and task scheduling

R provides RHadoop, a family of R packages, that acts as a wrapper for Hadoop streaming and allows execution of Hadoop jobs [1]. Some of the important packages from RHadoop include rhdfs and rmr. Rhdfs is used for handling the HDFS operations such as file storage and file manipulation whereas rmr is primarily responsible for the map-reduce function [1].

#### 4.2 Spark

Spark is another open-source distributed computing framework, however, unlike Hadoop, Spark is a memory based computing framework. Spark's in-memory processing capabilities often result in better performance as compared to Hadoop, especially when implementing iterative machine learning algorithms. The two key abstractions of Spark are:

- Resilient Distributed Datasets: Collection of fault-tolerant data items which can be operated in parallel.
- Directed Acyclic Graph (DAG): DAG is a set of vertices and edges where vertices represent the RDDs and edges represent the operations to be applied to the vertices. Spark's DAG engine optimizes the execution by breaking down a Spark job into complex multi-step data pipeline to be executed on the cluster.

R provides the package SparkR which is a light-weight frontend for using Apache Spark in R [1]. SparkR provides SparkContext which establishes a connection between the R program and the Spark cluster. Users can then use the RDD class provided by this package to explore the Spark API and interactively trigger jobs from

the R shell on to the Spark cluster. SparkR also provides distributed machine learning support through MLlib [1]. Thus, with the help of this package, we can implement a distributed version of the regression analysis.

## 5 REGRESSION TECHNIQUES FOR BIG DATA

Using the distributed version, we can efficiently apply different kinds of regression for big data analysis. We now provide an overview of some of these regression techniques:

### 5.1 Linear Regression

Linear regression is one of the most widely used regression techniques. It tries to model the relationship between a dependent variable and one or more independent variables using a best-fit straight line which is represented by the equation below:

$$y = \beta_0 + \beta_1 x$$

The above gives the regression line, where  $y$  represents the response variable and  $x$  represents the independent variable,  $\beta_0$  gives the intercept and  $\beta_1$  gives the slope of the regression line. Once the coefficients,  $\beta_0$  and  $\beta_1$  are estimated, the regression line can be used to predict values of the response variable,  $y$ , for a given value of the explanatory variable,  $x$ . Linear regression has applications in various fields such as weather forecasting, stock price prediction, etc.

### 5.2 Regularized Regression

We always aim to build a machine learning model that generalizes well on the unseen data. Similarly, for regression, we would want to find the coefficients for the independent variables such that the resulting regression line has minimum prediction error for, not only the training data but also for the unseen test data. In order to prevent over-fitting in regression, we use regularized regression techniques such as ridge regression and lasso regression. Ridge imposes  $L_2$  regularization whereas lasso imposes  $L_1$  regularization on the coefficients of the independent variables. Ridge regression does a better job in presence of multicollinearity, i.e., when two or more independent features are correlated. Lasso regression does a better job when the number of independent variables is large. This is because the penalty imposed by lasso can shrink some of the coefficients to zero. Thus, lasso regression also provides feature selection in case of a large number of features.

### 5.3 Logistic Regression

Unlike linear regression which is used to predict the continuous value of the response variable, logistic regression is used when the response variable has a binary outcome. Thus, logistic regression is used for classification problems. It is used to model the relationship between the categorical response variable and one or more independent variables by estimating the probabilities using a logistic function [1]. Logistic regression has applications in many fields such as medical and social sciences [1]. For instance, it can be used to predict whether the patient is suffering from a given disease based on the various attributes related to the patient like age, sex, body mass index, blood levels, etc. Another application for logistic regression is predicting whether a candidate will vote Democratic or Republican based on his age, sex, race, social economic status.

## 6 CONCLUSION

While many companies are now focusing on accumulating big data, efficient analysis of this large volume of data is significant. We explained various stages of traditional predictive modeling and showed how it can be extended for big data. We also discussed the distributed frameworks like Hadoop and Spark that can be used for processing big data. These frameworks are supported by some of the programming languages such as R. With the help of these frameworks and the machine learning libraries supported by R, we can implement various regression techniques for building a predictive model for big data. Thus, by partitioning big data, we can effectively build prediction models which can be useful in various fields like medical, social sciences, etc.

## ACKNOWLEDGMENTS

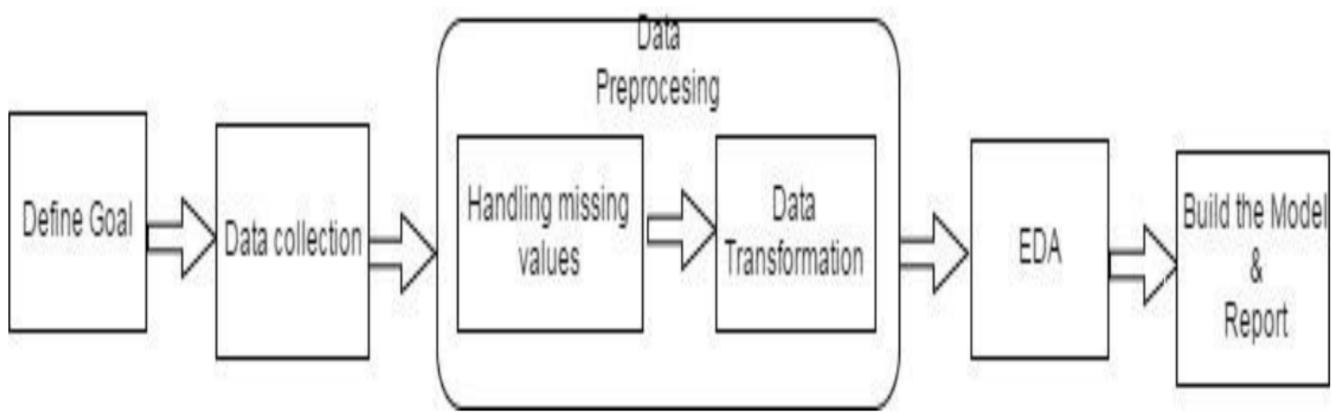
We would like to thank Dr. Gregor von Laszewski and the teaching assistants for their support and suggestions.

## REFERENCES

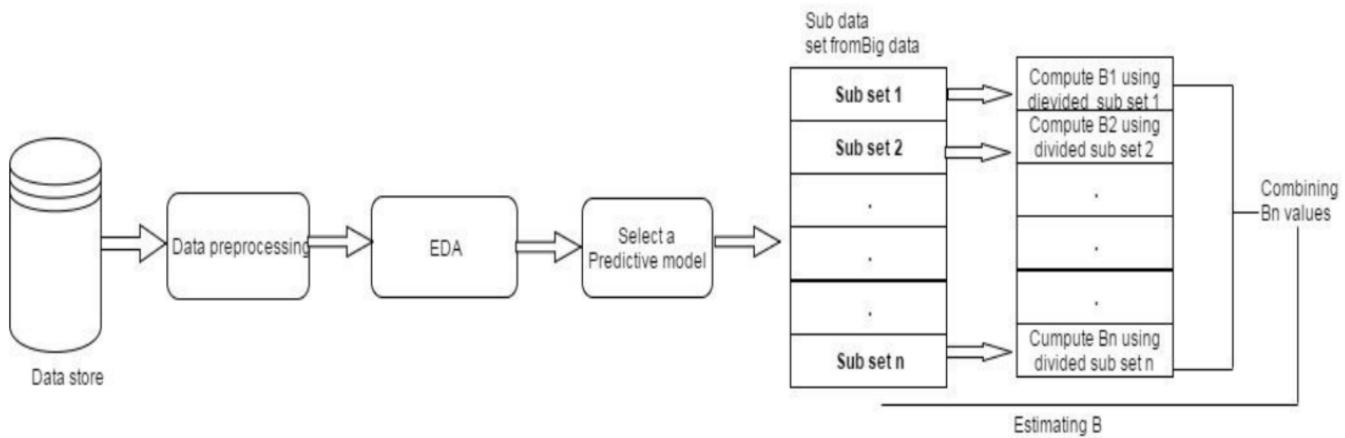
- [1] Ruizhu Huang and Weijia Xu. 2015. Performance evaluation of enabling logistic regression for big data with R. In *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, IEEE, Santa Clara, CA, USA, 2517–2524. <http://ieeexplore.ieee.org/document/7364048/>
- [2] Sunghee Jun, Seung-Joo Lee, and Jea-Bok Ryu. 2015. A Divided Regression Analysis for Big Data. *International Journal of Software Engineering and Its Applications* 9, 5 (2015), 21–32. [http://www.sersc.org/journals/IJSEIA/vol9\\_no5\\_2015/3.pdf](http://www.sersc.org/journals/IJSEIA/vol9_no5_2015/3.pdf)
- [3] K Saritha and Sajimon Abraham. 2017. Prediction with partitioning: Big data analytics using regression techniques. In *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on*. IEEE, IEEE, Thiruvananthapuram, India, India, 208–214. <http://ieeexplore.ieee.org/document/8076768/>
- [4] Chen Xu, Yongquan Zhang, Runze Li, and Xindong Wu. 2016. On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 3041–3052. <http://ieeexplore.ieee.org/document/7520638/>

#### LIST OF FIGURES

1	Steps in the predictive modeling process [3]	5
2	Architecture for partitioning Big Data [3]	5
3	A divide-and-conquer learning framework [4]	6



**Figure 1:** Steps in the predictive modeling process [3]



**Figure 2:** Architecture for partitioning Big Data [3]

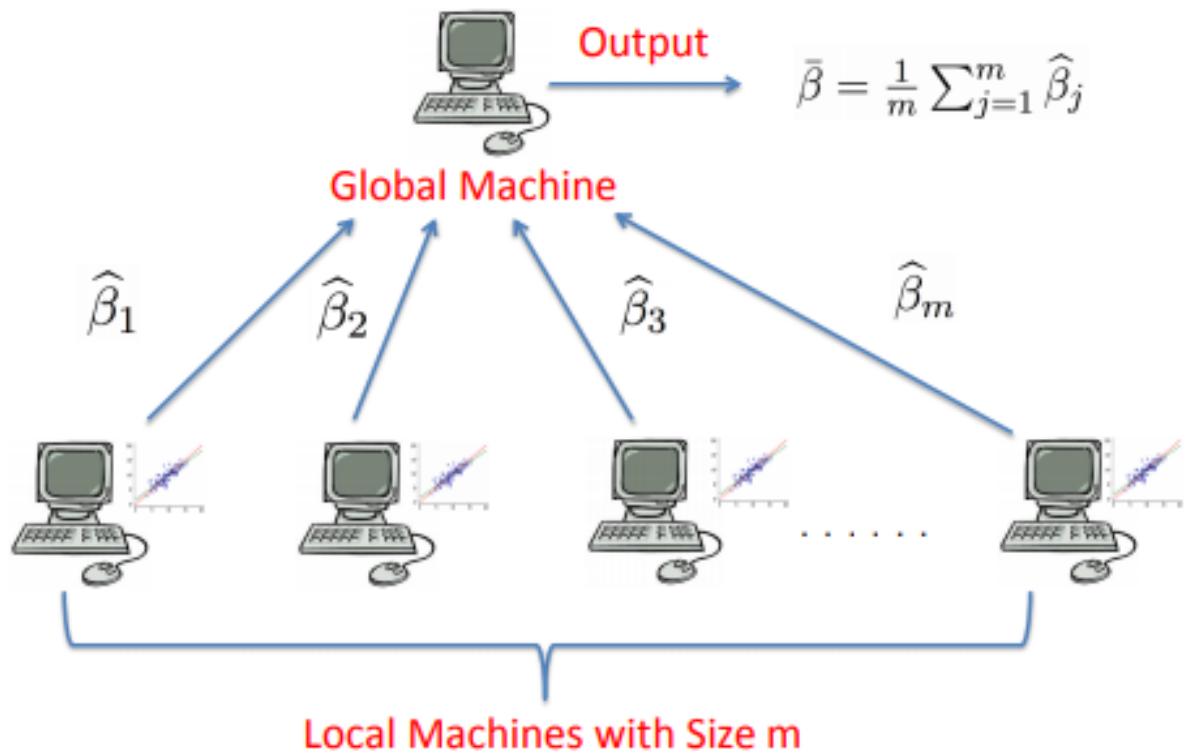


Figure 3: A divide-and-conquer learning framework [4]

```
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
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-11-06 17.35.33] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.6s.
./README.yml
21:81      error      line too long (177 > 80 characters)  (line-length)
```

---

```
Compliance Report
```

---

```
name: Chandwani, Nisha
hid: 203
paper1: 100% Nov 08 2017
paper2: 100% Nov 06 2017
project: not started
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
```

```
wc 203 paper2 6 565 content.tex  
wc 203 paper2 6 2406 report.pdf  
wc 203 paper2 6 145 report.bib
```

```
find "
```

---

```
102: Do not use "these quotes" but use these ``these quotes''.  
passed: False
```

```
find footnote
```

---

```
112: \footnote{do not use footnotes}.
```

```
passed: False
```

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

---

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

```
passed: True
```

```
floats
```

---

```
62: In Figure \ref{f:fly} we show a fly. Please note that because we  
use  
68: \begin{figure}[!ht]  
69: \centering\includegraphics[width=\columnwidth]{images/fly.pdf}  
70: \caption{Example caption}\label{f:fly}  
85: or generate them by hand while using the provided template in  
Table\ref{t:mytable}. Not ethat  
88: \begin{table}[htb]  
91: \label{t:mytable}
```

```
figures 1  
tables 1
```

```
includegraphics 1
labels 2
refs 2
floats 2

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

#### Label/ref check

```
105: 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 change the number to a smaller fraction
```

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors

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

```
bibtex_empty_fields
```

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Applications in Virtual Assistants

Jiaan Wang

Indiana University Bloomington

3209 E 10 St

Bloomington, IN 47408

jervwang@indiana.edu

## ABSTRACT

This paper provides

## KEYWORDS

i523, HID233, Big data, Virtual Assistants, Artificial intelligence

## 1 INTRODUCTION

"Big data is a technical buzzword that is recently grown quite popular in tech circles. It is a term that generally refers to the collection and storage of a giant horde of data, a horde so large that only supercomputers can chew through it. We are talking data at the petabyte scale" [7].

"In the past, all this data was impossible to sort through, but with each passing year better algorithms, coupled with increasingly powerful supercomputers, have allowed governments and corporations to connect the dots and find patterns in all this data. These patterns then allow organizations to better execute three important functions: Control increasingly complex systems (like city utilities and corporate logistics), improve existing systems (general government services and flight path planning), and predict the future (weather and financial forecasting)" [7].

"As you can imagine, the applications for big data are immense. It will allow organizations of all kinds to make better decisions about the services and systems they manage. But big data will also play a big role in helping texts, your emails, your social posts, your web browsing and search history, the work you perform, who you call, where you go and how you travel, what home appliances you use and when, how you exercise, what you watch and listen to, even how you sleep! On any given day, the modern individual is generating huge amounts of data, even if he or she lives the simplest of lives. This is big data on a little scale" [7].

"Search engines are undoubtedly a staple in our everyday lives - for most of us, we rely on the search giant, Google, which provides us with tailored search results to many questions throughout the day. But many technology and data companies are realizing that the next generation of search lies in vertical, or topic-specific, search. Rather than solving large, more general problems, vertical search tackles more specific and precise queries. Vertical search is even beginning to emerge within industries such as travel, entertainment, fashion and more. This dynamic presents an opportunity for other companies to surpass Google in industry-specific verticals" [4].

"As an example, Zite, a news recommendation smart-phone app, offers the end-user recommendations of what to read based on preferences and uses artificial intelligence to learn the behaviors and preferences to create and continuously improve these recommendations. This creates intelligent search recommendations that

are much more vertical and specific than what the end-user would experience with Google" [4].

"Truly intelligent vertical search engines utilize text and image classification coupled with other AI algorithms and big data analytics to gain detailed knowledge about both users and content in verticals or applications. It will be these companies can gain a competitive advantage over giants like Google" [4].

"Retailers in particular have the opportunity to monopolize on next-generation virtual assistants, as they offer the opportunity to directly relate to shoppers and create truly loyal customers. The ideal virtual assistant of the future will help fill in the gaps that currently exist between a personalized experience that in-store shoppers are used to, and current, less helpful, online and mobile shopping applications. Some virtual assistants have already begun to use geo-targeting technologies to localize experience, but the next generation will focus on a direct interaction with the customer that will create a seamless customer experience if handling every step of the purchase cycle. We will continue to see more of these emerge as omni-channel retail becomes a larger part of our everyday lives, beginning with mobile virtual assistant apps" [4].

"You are probably familiar with Amazon's recommendations, a first-generation example of online product suggestions, as Amazon uses collaborative filtering. This method, however, is not capable of providing suggestions that present what the user actually wants. Collaborative filtering attempts to filter products driven by taste - for example, if you buy a pink shirt, it may then suggest you buy that same shirt in red" [4].

"Product advice powered by AI offer the next-generation of recommendations. By extracting data from multiple sources including your location and amount of time spent on the site, retailers will build the knowledge to predict their customers' preferences and needs. Machine learning will begin to allow retailers to process this data and generate a deep knowledge of not only their products but their users, and even more importantly, their preferences and behaviors. Better recommendations can be built by classification of products and multiple recognition and data enhancement methods - laying the groundwork for retailers to establish meaningful relationships with their customers by recommending them truly relevant products" [4].

"As more apps and technology begin to incorporate AI, the more able they will become to predict behaviors of the end-user. You will begin to see more systems that learn your behaviors and are able to provide you with a much more personalized, seamless user-experience, as AI will continue to open up new doors to incorporate more abstract and difficult data sources" [4].

"How does machine learning work in virtual assistants? Firstly, it is a combination of several algorithms as multiple issues need to be resolved. Initially, a semantic token parser will be required. This

then needs an expert database (logical networks such as Prolog, object based databases, functional representation) of semantic representation. After that, you will implement the learning structures based on existing knowledge or by programming it. This serves the pattern matching algorithm. The next step is finding the valid answer in the network, which requires a different algorithm (graph). Then, you'd need an interface algorithm for user interaction (IRC channel, web, web socket, general socket, API). You'll also need a semantic generator - an algorithm that is generating from the found solution a grammatically and syntactically correct natural language way to represent the answer in a human readable way" [6].

"Thanks to the sudden acceleration of artificial intelligence and advancements in speech recognition and big-data storage, the technology behind virtual assistants is rapidly spreading from phones to cars and homes, and the truly useful helper is approaching fast. The four giants are fighting for the biggest share of a market expected to grow to 12 billion dollars by 2024" [1].

"The ultimate goal is our own personal genie in a bottle that awakens with a word or touch to liberate us from all our mundane tasks, organize our days and nights, and free us from the stress of lives that have become so terribly busy. But that's not going to happen quite yet" [1].

"The industry stands at a critical moment, because the first highly effective help-bot to get a foothold in a consumer's home, phone or car will likely stay, creating a barrier to competitors" [1].

"In order for a virtual helpmate to run your life, it needs to engage with the providers of all the services you rely on, from your calendar app to your Uber ride. Those providers must either partner with the company operating the assistant, or design their app to integrate with the assistant. So Spotify will stream music upon request via Alexa, and Honeywell's smart-home thermostat, via Assistant, will bump up the temperature 15 minutes before Grandma's expected arrival" [1].

## 2 CURRENT APPLICATIONS IN THE FIELD

"Virtual assistants rely on a number of different technologies. One involves speech recognition. Google says it has reduced the error rate for recognizing words in its own mobile app to less than 8 per cent - a level at which it says the service has become a practical alternative to entering text. Apple has also made considerable headway in overcoming the early disappointment by adding more capabilities to Siri and helping users better understand what it can and cannot do" [8].

"A second front has involved the predictive technology for anticipating what a user will want to know next. This draws on contextual data - aspects such as the location and time of day - as well as personal information. Knowing what other people have found useful is also valuable" [8].

"Future VAs will use all of this data to better understand you with the goal of helping you accomplish your daily tasks more effectively. In fact, you may have already used early versions of VAs: Google Now, Apple's Siri, or Microsoft's Cortana" [7].

"Each of these companies have a range of services or apps to help you collect, store, and use a treasure trove of personal data. Take Google for example. Creating a single Google account gives

you access to its large ecosystem of free services - search, email, storage, maps, images, calendar, music and more - that are accessible from any web-enabled device. Every action you take on these services (thousands per day) is recorded and stored in a *personal cloud* inside Google's server farms. With enough use, Google begins to understand your preferences and habits with the end goal of using *anticipatory systems* to provide you with information and services you need, when you need it, before you even think to ask for it" [7].

"For example, Apple users generally use Apple desktops or laptops at home and Apple phones outdoors, all while using Apple apps and software in between. With all these Apple devices and software connected and working together within the Apple ecosystem, it shouldn't come as a surprise that Apple users will likely end up using Apple's VA: A future, beefed up version of Siri. Non-Apple users, however, will see more competition for their business" [7].

"Google already has a sizable advantage in the machine learning field. Because of their globally dominant search engine, popular ecosystem of cloud-based services like Chrome, Gmail, and Google Docs, and Android, Google has access to over 1.5 billion smartphone users. This is why heavy Google and Android users will likely choose a future version of Google's VA system, Google Now, to power their lives" [7].

"While seen as an underdog due to its near non-existent market share in the smart-phone market, Microsoft's operating system, Windows, is still the dominant operating system among personal desktops and laptops. With its 2015 roll-out of Windows 10, billions of Windows users around the world will be introduced to Microsoft's VA, Cortana. Active Windows users will then have an incentive to download Cortana into their iOS or Android phones to ensure everything they do within the Windows ecosystem gets shared with their smart-phones on the go" [7].

"Cortana and Google Assistant are getting smarter, with contextual reminders and recommendations geared toward optimizing productivity along with fascinating innovations incorporating computer vision and other machine learning algorithms. Alexa is building out a diverse ecosystem of third-party skills, and Google and Microsoft have followed suit" [5].

"Alexa Skills Kit, Cortana Skills, and Actions on Google give companies and developers the tools to apply the voice tech to everything from email marketing and e-commerce to expense tracking and fleet management. These business applications and use cases are only what we've seen so far. PCMag spoke to execs from Amazon, Google, and Microsoft to understand their virtual assistant vision, how the tech is evolving, and what these companies believe businesses can do with voice-capable AI helpers" [5].

"The value of a virtual assistant is having it there wherever you are, giving you the tailored information you need sometimes before you even know you need it. Microsoft, more than the other tech giants building out this tech, has deeper roots in business software and productivity. Cortana is enabled across a number of Microsoft's apps and services - from Microsoft Power BI to Skype to provide immediate contextual responses to business queries without leaving the app you are in" [5].

"The simplest way Cortana does this is through reminders. Scheduling, reminders, and lists are a top-of-mind business use case for virtual assistants. Jones talked about using Cortana in a touchscreen

device such as the Microsoft Surface Pro. Smart Sticky Notes in the Windows 10 Anniversary Update let you write something like, *Call my boss at 3pm*, either by typing or by writing a note with the stylus as part of Windows Ink. Cortana will then add that reminder to keep track of the task” [5].

“Microsoft is also working with Wunderlist to integrate Cortana and sync lists across devices. This is all part of a more proactive strategy, using both contextual data and location-based reminders to help users manage their emails, schedule, and day-to-day commitments. Microsoft is looking to expand this even further to dynamically create Cortana to-do lists and surface information based on data throughout Office 365. Cortana is already fully integrated into the Microsoft Edge browser and can search for documents or people across apps such as and SharePoint” [5].

“The Windows 10 Creators Update also integrated Cortana with Microsoft Azure Active Directory (AAD) to bring the AI capabilities to enterprise users who may not have had access to it before. These kinds of integration also extend to Power BI, which lets you pull Cortana data into business intelligence queries and reports. That’s not to be confused with the Cortana Intelligence Suite, a separate enterprise offering that builds machine learning and predictive analytics into business apps” [5].

“Beyond that, Jones said the Cortana team is working with Microsoft Research through projects such as *Calendar.help* to automate processes like scheduling meetings with contacts outside your organization. The team is also working with the Microsoft IT Division development team to create experiences specific to Cortana that pull in a range of apps and contextual data” [5].

“The more tasks you teach and program an AI to perform, the more it will be able to do. In this respect, virtual assistants have something in common with the deep learning process by which ML algorithms and neural networks are trained on massive data sets. Training virtual assistants to perform specific business tasks is easier; all you have to do is open up the ecosystem to third-party skills development” [5].

“Amazon is the standard-bearer in this regard. The Alexa Skills Kit has been available since 2015 and lets companies and developers apply Alexa to whatever business environment or process they desire. As a result, there is already a wide selection of available business skills that companies can simply enable and start using - and that ecosystem is growing” [5].

“Google and Microsoft have followed Amazon’s lead on this front with Actions on Google and Cortana Skills, respectively. These tool-kits let you build particular skills but they’re also evolving to incorporate natural language processing and features such as proactive suggestions to recommend a skill to users in the right context” [5].

“Interoperability aside, the fact is, this space is still only a few years old. Amazon launched The Alexa Fund last year to spur innovation in the space, committing to invest up to 100 million dollars in venture capital funding to both start-ups and established brands pushing the boundaries of what voice and virtual assistant tech can do. Google and Microsoft are both heavily invested in continued research as well” [5].

“Now, perhaps unsurprisingly, Facebook pops up again. In the last chapter for this series, we mentioned how Facebook will likely

enter the search engine market, competing against Google’s fact-focused semantic search engine with a sentiment-focused semantic search engine. Well, in the field of VAs, Facebook can also make a big splash” [7].

“Facebook knows more about your friends and your relationships with them than Google, Apple, and Microsoft together ever will. Initially built to compliment your primary Google, Apple, or Microsoft VA, Facebook’s VA will tap into your social network graph to help you manage and even improve your social life. It will do this by encouraging and scheduling more frequent and engaging virtual and face-to-face interactions with your friend network” [7].

“Over time, it is not hard to imagine Facebook’s VA knowing enough about your personality and social habits to even join your circle of true friends as a distinct virtual person, one with its own personality and interests that reflect your own” [7].

“Today, the aid these virtual assistants provide remains limited. Most users of Google Home and Amazon Echo devices - which host Assistant and Alexa, respectively - stream music, play audio books, and control smart-home devices, according to surveys by San Francisco analytics firm VoiceLabs” [1].

“Still, the virtual agent’s foundation in AI means the more it learns about a user’s preferences and behaviors, the better job it can do. So while experts predict a handful of firms will dominate in this field - most agree Apple, Google and Amazon will be major players, with Microsoft in a lesser role - they are split on whether consumers will be served best by one bot, or more” [1].

“People want one assistant, they do not want two. You want one assistant, to be very readily available wherever you are. However, the various assistants will likely end up somewhat specialized in their expertise, with Google Assistant, for example, excelling in providing knowledge and managing schedules, and Microsoft Cortana leading on gaming. In a few years, many people will use two or three different assistants” [1].

“For all the major players, virtual assistants provide important data that fuels the AI that powers and improves them, making both the assistants and the products they live in ever more marketable. For Amazon, Alexa is an enthusiastic purchasing agent for the e-commerce that drives the firm. For Google, Assistant is a turbocharged vacuum for the data the company collects to sell ads targeted directly at users” [1].

“So far, both Google and Amazon have focused largely on home-based assistants. Google’s new Pixel phones host Assistant, but it has an uphill battle because Apple has far more phones equipped with Siri on the market” [1].

“The popularity of Amazon Echo and Alexa notwithstanding - the company has sold more than 8 million Echo devices since rolling them out in late 2014, according to Consumer Intelligence Research Partners - most people want their virtual assistant on their phones” [1].

### 3 THE FUTURE OF VIRTUAL ASSISTANTS

“Putting these robo-helpers into cars’ on-board systems has become a priority for major firms, including Microsoft, which seeks to extend the reach of its PC-based Cortana through the *connected-vehicle* platform it announced this year” [1].

"In January, Nissan announced it would integrate Microsoft's platform into its cars. Siri already can be used in a car via a phone or Apple's CarPlay system, or in cars sold with Siri integration built in. Hyundai is bringing Alexa and Google Assistant into some of its cars so, for example, an owner could start their car from their living room" [1].

"Slack CEO Stewart Butterfield has an audacious goal: Turning his messaging and collaboration platform into an uber virtual assistant capable of searching every enterprise application to deliver employees pertinent information. And if Slack succeeds, it could seal the timeless black hole of wasted productivity enterprise search and other tools have failed to close" [2].

"The real potential comes in the form of intelligent virtual assistants, known as chatbots. Slack this year introduced a platform and development kit that allows third-party developers to build bots designed to make tedious tasks such as managing expenses, tracking projects or ordering tacos more efficient. If developers create enough bots, employees won't have to switch out of Slack to access apps in browser windows" [2].

"For example, suppose that you wanted to know who someone's boss was, or what a business unit's revenue was for a quarter. You could ask around or sift through a corporate directory laden with an enterprise search system. But what if you could just ask a bot, which could retrieve the answer almost instantly? You can build institutional knowledge and ask that of a bot instead of a human and it saves people a lot of time and offloads a lot of noise" [2].

"Slack could eventually train bots to recognize when conversations are going on to too long without a resolution in sight and recommend that the team members conduct a face-to-face or virtual meeting - and then it would schedule it for them. Slack has a position as an interface that makes sense if messaging is the way you want to interact with the bots" [2].

"Slack is building an enterprise version that will include many of the necessary attributes CIOs have come to demand from productivity and collaboration tools, including the capability to provision and de-provision users and have fine-grained control and policy setting over channels through a single web dashboard. The software, currently in testing with half a dozen businesses, will feature metrics about consumption and other analytics. When it does launch, platforms from Microsoft, IBM and the just launched Facebook Workplace platform will be waiting" [2].

"A new crop of virtual assistants is on the way, led by Amy, Shae and Otto. Each in its own way represents the future of virtual assistants. One is in public beta, one is in private beta and one is a hardware prototype, but they are coming soon, and they collectively reveal how much better virtual assistants can be" [3].

"Amy does one thing really well: scheduling your meetings. Amy is the creation of a New York startup called *x.ai*. For now, Amy lives on the other side of an email address: *amy@x.ai*. Company intends to put Amy on other platforms, such as Slack and other group-chat apps, Amazon Echo and more, and that the platform shouldn't matter" [3].

"You simply cc: Amy's email address on your communication about the scheduling of any meeting, and Amy takes over. Amy is *invisible software* - there is no app to install, no website to interact with" [3].

"Amy is adept with natural language processing, which means you can use everyday language. For example, you might send an email to a colleague, copying Amy, and say: *Hey, let's get together next week* or *Grab a bite next week?* or *We should connect.* Amy will then take action, introduce herself to the other person and, based on your calendar and preferences, suggest a time to meet" [3].

"Amy is interactive. If the person you want to meet with gets back to Amy with restrictions or additional suggestions, Amy handles all the back-and-forth that often attends the hunt for a mutually agreeable meeting time. If you want to know how it's all going, you can send an email to Amy and ask how the meeting with so-and-so is going and Amy will reply with the current status. If anyone wants to reschedule later, Amy handles that in the same way" [3].

"Best of all, Amy does the heavy lifting when you need to reschedule. Let's say you decide to take a last-minute vacation. Just send an email to Amy and say: *Clear my schedule for next week.* If you have got 10 meetings scheduled, Amy will reach out to all 10 people to reschedule and will update your calendar" [3].

"Amy represents the future of virtual assistants for two reasons. First, it's a specialist agent, doing one thing very well. Second, Amy is believably human. Within the confines of email conversations on the subject of scheduling meetings, Amy passes the Turing test" [3].

"Shae helps you get healthy by guiding and informing you about healthy living all day, every day. The company behind Shae, *Personal Health 360* or *PH 360* throws around some big numbers. It claims Shae uses some 500 algorithms fed by more than 10,000 data points to provide very specific help for users" [3].

"That is a lot of data, and it comes from unexpected places. For example, family history is taken into account, individual body type and environmental factors like the weather and pollen count. Much of the health data Shae uses comes from a personalized phenotype questionnaire that each user fills out" [3].

"Shae additionally accesses both your calendar and bio-metrics as detected from a monitoring device like the Apple Watch to figure out what your mood might be. When it detects signs of stress such as an elevated heart rate, the app pops up a dialog to ask you if you are feeling stressed" [3].

"Like Google Now, Shae takes the initiative to give you information, updates and advice, telling you what to eat and when to exercise, and keeping tabs on changing health data, such as your weight, body mass index, lean muscle mass and other body measurements. Shae even helps you plan vacations based on your personal profile and circadian rhythm" [3].

"Over time, we will learn if users are thrilled with the Shae assistant. Whether Shae succeeds or fails, the service represents the future of virtual assistants because of its extreme personalization and the eclectic nature of the data, integrating family history, personal health details, health knowledge, environmental data and more - and for its preemptive advocacy of habits that benefit the user" [3].

"Following the success of Amazon's Echo device, Samsung unveiled its own virtual assistant home appliance called Otto in April of this year. Like the Amazon Echo, which is possessed by a virtual assistant named Alexa, the Otto is an Internet-connected speaker and microphone that interacts with you via spoken conversations

and can control home appliances like lights. Otto can answer questions, order products, and play music and pod-casts on command. But unlike the Echo, Otto is also an HD security camera that can stream video live to your phone or computer. The device has a kind of head that can turn, pivot and swivel to let you look around the room remotely. Otto can also recognize faces - and even has a rudimentary face of its own, displayed on a screen” [3].

“It is based on Samsung’s ARTIK IoT platform, which Samsung recently unveiled developer tools for. Otto represents the future of virtual assistants not only because it is a physical home appliance, but also because it uses facial recognition. That means different members of the family could each have their own set of preferences and personal details, calendars and accounts - and Otto and other future appliances could base their interactions on the knowledge of the person they are talking to” [3].

“Shae, Amy and Otto together represent the future of virtual assistants, which will be specialized, personalized, thorough, pre-emptive, highly intelligent and optionally available in the form of dedicated physical appliances. These three virtual assistants already suggest just how helpful and, well, human our technology will become” [3].

## 4 CONCLUSION

Put here an conclusion.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Ethan Baron. 2017. One bot to rule them all? Not likely, with Apple, Google, Amazon and Microsoft virtual assistants. Web Page. (Feb. 2017). <http://www.mercurynews.com/2017/02/06/one-bot-to-rule-them-all-not-likely-with-apple-google-amazon-and-microsoft-virtual-assistants/> HID: 233, Accessed: 2017-10-24.
- [2] Clint Boulton. 2016. Slack CEO describes ‘Holy Grail’ of virtual assistants. Web Page. (Oct. 2016). <https://www.cio.com/article/3131536/collaboration/slack-ceo-describes-holy-grail-of-virtual-assistants.html> HID: 233, Accessed: 2017-10-24.
- [3] Mike Elgan. 2016. These three virtual assistants point the way to the future. Web Page. (June 2016). <https://www.computerworld.com/article/3078829/artificial-intelligence/these-three-virtual-assistants-point-the-way-to-the-future.html> HID: 233, Accessed: 2017-10-18.
- [4] Lars Hard. 2014. The Disruptive Potential of Artificial Intelligence Applications. Web Page. (Jan. 2014). <http://data-informed.com/disruptive-potential-artificial-intelligence-applications/> HID: 233, Accessed: 2017-10-18.
- [5] Rob Marvin. 2017. What Are Virtual Assistants and What Can You Do With Them? Web Page. (June 2017). <https://www.pcmag.com/article/354371/what-are-virtual-assistants-and-what-can-you-do-with-them> HID: 233, Accessed: 2017-10-24.
- [6] Anubhav Srivastava. 2016. Why the virtual assistants market is on the upswing? Web Page. (July 2016). <http://thinkbigdata.in/virtual-assistants-market-upswing/> HID: 233, Accessed: 2017-10-24.
- [7] David Tal. 2015. Forecast – Rise of the big data-powered virtual assistants: Future of the Internet P3. Web page. (Nov. 2015). <http://www.quantumrun.com/prediction/rise-big-data-powered-virtual-assistants-future-internet-p3> HID: 233, Accessed: 2017-10-18.
- [8] Richard Waters. 2015. Artificial intelligence: A virtual assistant for life. Web page. (Feb. 2015). <https://www.ft.com/content/4f2f97ea-b8ec-11e4-b8e6-00144feab7de?mhq5j=e5> HID: 233, Accessed: 2017-10-18.

## bibtex report

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

bibtext \_ label error

bibtext space label error

## bibtext comma label error

# latex report

[2017-11-06 17.36.33] pdflatex report.tex

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

```
name: Wang, Jiaan
hid: 233
paper1: Nov 03 17 100%
paper2: Nov 10 17 80%
project: 10%
```

```
yamlcheck
-----
```

```
wordcount
-----
```

```
5
wc 233 paper2 5 4449 content.tex
wc 233 paper2 5 4422 report.pdf
wc 233 paper2 5 298 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
-----
```

```
passed: True
```

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

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

```
passed: True
```

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

```
figures 0
tables 0
```

```
includegraphics 0
labels 0
refs 0
floats 0

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

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

# Hadoop and MongoDB in support of Big Data Applications and Analytics

Sushant Athaley  
Indiana University  
sathaley@iu.edu

## ABSTRACT

Big data processing is beyond the capability of traditional tools. It requires specialized tools to handle the volume, velocity, and variety of big data. We explore Hadoop and MongoDB technically as a tool and how they provide support/help in big data analytics.

## KEYWORDS

i523, hid302, big data, Hadoop, MongoDB, HDFS, MapReduce

## 1 INTRODUCTION

The emergence of big data challenges also gave rise to the various technologies which can be used to solve big data problem. Typically to solve big data, we need to consider two types of technologies, data capture and storage, and data analysis. We evaluate capabilities of two popular technologies Hadoop and MongoDB to understand their features and power to solve big data problem. We get started with the section *Big Data* which captures big data definition and characteristics. Section *Hadoop* provides an introduction to the technology and subsections *Hadoop Common*, *HDFS*, *YARN*, *MapReduce* explores technology in detail. Section *Big Data Support* captures how Hadoop supports big data analytics along with the real-life examples. We then explore MongoDB through section *MongoDB* which further drill down to the technology using sections *Architecture*, *Data Model*, *Data Management*, *Data Visualization*, and *Security*. Section *Big Data Support* captures how MongoDB supports big data analytics along with the real-life examples. Section *Power of Two* captures how both technologies can be used together to solve big problems. Section *Conclusion* concludes the study.

## 2 BIG DATA

Big Data is defined in lot many different ways but one of the interesting ways it has been defined is in terms of three V's which are Volume, Velocity, and Variety. Big data is generated in great *volume* typically in the gigabyte or more which makes data processing difficult. Data *velocity* has been increased due to the real-time data streaming from various applications like social media or different type of sensors recording data continuously. Big data comes in *variety* of format like structured or unstructured data. Data varies in various format like text, pictures, audio, videos, 3D, social media and so on. These big data characteristics pose challenges in terms of overall data lifecycle management. Some of the examples of big data usage are the recommendation service, predictive analytics, data analytics, pattern identification, and machine learning. Traditional systems are good for small or medium data processing but unable to provide support for the big data. Big data need specialized technologies and tools to handle its characteristics. The technologies which can solve big data problem should have capabilities

like distributed computing system, massively parallel processing, NoSQL, and analytical database [1, Ch. 1, p. 4]. Can Hadoop or MongoDB be those technologies who can provide that support?

## 3 HADOOP

Apache foundation describes Hadoop as “The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures” [2]. In other words, Hadoop provides a framework to store data in the distributed manner and provides the capability to run data analysis in the distributed way.

“Currently Hadoop project includes following modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets” [2].

### 3.1 Hadoop Common

Hadoop Common is the collection of the utilities to support the Hadoop modules. This is the core package which provides essential and basic service of the framework.

### 3.2 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is the default distributed file system provided by the Hadoop. HDFS serves as storage mechanism in the Hadoop framework. HDFS specifically designed to process large data set and run on low-cost hardware. It is highly fault-tolerant which contains the mechanism for quick fault detection and auto recovery. HDFS is designed to port across heterogeneous hardware and software platform. It does data computation on the same node instead of moving data to the server which is faster as well as avoid network congestion. It provides scalability by adding or removing nodes in the HDFS cluster and can support hundreds of nodes in single cluster [4]. Figure 1 shows HDFS architecture.

[Figure 1 about here.]

HDFS is based on master/slave architecture where NameNode is the master server and DataNodes are the slave nodes. There can be only one NameNode server which manages file system namespace and all read-write requests. NameNode doesn't store any data but contains all the meta-data about files and DataNodes. DataNode contains actual data and they can be multiple in numbers usually one per node. DataNodes are responsible for the create, delete, replicate of the data blocks on the node as per the instruction by the NameNode. DataNode also sends block-report to NameNode which has a list of all blocks on the DataNode. DataNode sends the heartbeat message to NameNode which helps in identifying the failure nodes. If the heartbeat is not received by NameNode in specified interval then that DataNode is marked as dead and NameNode usage different DataNode. Figure 2 and 3 depicts read and write in HDFS respectively.

[Figure 2 about here.]

[Figure 3 about here.]

### 3.3 Hadoop YARN

Hadoop YARN (Yet Another Resource Negotiator) provides cluster resource management which helps in running multiple distributed application in Hadoop. YARN consists of 3 components *ResourceManager (RM)*, *NodeManager (NM)* and *ApplicationMaster (AM)*. ResourceManager is the master process which manages resources across the nodes. NodeManager is responsible for the container and provides resource usage to the ResourceManager. ApplicationMaster is responsible for getting resources from ResourceManager and work with NodeManager to execute the task [3]. YARN makes it possible to run different applications on Hadoop platform which makes it scalable and integrable [1, Ch. 3, p. 65]. Figure 4 shows YARN architecture.

[Figure 4 about here.]

### 3.4 Hadoop MapReduce

Hadoop MapReduce is a framework which provides the capability to process the vast amount of data in a distributed manner. Processing is done in parallel on various nodes utilizing local machine processor and memory which results in high computation power. The framework provides fault tolerance along with supporting large clusters usually thousands of nodes. Typical framework processing is to split input data into independent chunks and then processed by *map* tasks in parallel. Sort the output of the map task and then provide that as input to *reduce* task for aggregate processing. Two important classes in this framework under package org.apache.hadoop.mapreduce are Mapper and Reducer. They respectively provide map and reduce method to process the data.

Figure 5 shows MapReduce process using wordcount example. Each line in the input file is passed to individual mapper class. Mapper class parses the line and sets count for the word. Sort and shuffle consolidate the data and sends it to the reducer. Reducer performs the final word count and provides the output.

[Figure 5 about here.]

### 3.5 Big Data Support

Big data problem solution requires tools which can process the huge amount of data with high computation power. Hadoop provides this capability by processing data in the distributed environment in big clusters using MapReduce and also provides distributed storage system as HDFS. HDFS can be configured in a cluster of hundreds of nodes and can typically store the file in size of gigabytes or terabytes [4]. Using CPU and memory of local nodes delivers great computation power. Hadoop also provides high tolerance to the faults and scalability by adding nodes and integration with various technologies. Being open source and configurable on commodity hardware, Hadoop is cost effective and can be used by small industries as well for their big data solution. Hadoop's capability to process large-scale data in parallel within distributed environment makes it one of the best tool for Big Data processing. Hadoop is supported by various sub-projects which together forms Hadoop Ecosystem. Different applications can be integrated with Hadoop depending on the big data problem need to be solved. Figure 6 illustrates Hadoop ecosystem by various layers.

[Figure 6 about here.]

Yahoo has one of the biggest Hadoop clusters. It has more than 100,000 CPUs in 40,000 computers running Hadoop. Their biggest cluster is of 4500 nodes. Yahoo is using Hadoop in research of Ad systems and web search and also used to do scaling tests to support the development of Apache Hadoop on larger clusters [5].

Facebook uses Apache Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning. They have 2 major Hadoop cluster, a 1100-machine cluster with 8800 cores and about 12 PB raw storage and a 300-machine cluster with 2400 cores and about 3 PB raw storage. Each node has 8 cores and 12 TB of storage [5].

Ebay has 532 nodes Hadoop cluster. They are heavy user of Mapreduce, Apache Pig, Apachae Hive and Apache Hbase. They are using Hadoop for search optimization and research [5].

## 4 MONGODB

The rise of Big Data started posing challenges on how data can be stored and processed. The inability of the traditional relational database to scale to big data volume and variety gave rose to the NoSQL databases. Based on the concept of one size doesn't fit all, NoSQL implementation comes in 4 different flavors, namely Column/Column Family, document, key-value and graph. MongoDB is one of the popular implementation of the document type NoSQL [6].

### 4.1 Architecture

MongoDB architecture blends best of relational and NoSQL technologies. It is enriched from relational database learning and incorporated new NoSQL features. Figure 7 shows MongoDB architectural consideration.

[Figure 7 about here.]

MongoDB provides powerful query language, indexing capability, strong read-write consistency, the capability to integrate with other technologies which they borrowed from the relational database. It

also implemented NoSQL features like flexible data model (schema-less), scalability by sharding or partitioning, and provides high availability by running across the nodes and replication mechanism. MongoDB also provides a multi-model architecture which supports four different storage engines and flexibility to mix and match those storage engines to store data in single MongoDB deployment [7].

## 4.2 Data Model

MongoDB stores data as BSON (Binary JSON) document object which is an extension of JSON and includes additional data type such as int, long, date, floating point, and decimal128. Documents are stored in the collection which is similar to row and table in relational database. The document can vary in structure and usually contains entire object details in the same document. This provides desired structure flexibility in terms of storing data which is not present in the relational database. The document can miss some fields and can be added to the document at any given point time without impacting other documents. Unlike other NoSQL databases, MongoDB provides data validation at the database level. Checks can be enforced at the database level to validate document structure, data types, data ranges and mandatory values [7].

## 4.3 Data Management

MongoDB has the capability of horizontal scaling which is referred as Auto-Sharding. In case of data increase, MongoDB distributes data across multiple physical partitions called shards automatically without impacting the application.

MongoDB is ACID compliant at the document level, the entire document is updated in one transaction or error is thrown.

MongoDB maintains multiple copies of data replica using native replication method. In case of failure, primary replica takes over giving high availability. This is done without impacting the application and is fully automated.

Ops Manager gives developers, administrators and operations teams monitoring capabilities into the MongoDB service. Featuring charts, custom dashboards, and automated alerting, Ops Manager tracks 100+ key database and systems health metrics including operations counters, memory and CPU utilization, replication status, open connections, queues and any node status.

Disaster recovery is provided using backup and restore mechanism. Backups are taken just a few seconds behind the operational system [7].

## 4.4 Data Visualization

MongoDB provides visualization capabilities in MongoDB Enterprise Advanced version using MongoDB Connector for BI. The tool provides the capability to analyze unstructured MongoDB data along with structured SQL database data [7].

## 4.5 Security

Security is a growing concern and MongoDB addresses it by providing extensive security features in MongoDB Enterprise Advanced. *Authentication* provides integration with external security mechanism like LDAP, Windows Active Directory, Kerberos etc. *Authorization* can be provided using the user-defined role to access the data also certain data can be masked by using a view. *Audit*

capability provides tracking of any command executed on the database. *Encryption* can be used to encrypt the data on the disk, on the network or in backup [7].

## 4.6 Big Data Support

MongoDB is a perfect fit for solving big data problem in terms of database storage. It is capable of handling big data volume and velocity using sharding which scales horizontally. It handles big data variety by providing schema-less data storage. Structured or unstructured, any type of data can be stored in MongoDB. It provides cost benefit as it can be installed on low-cost hardware as well as by cutting down on the development time of the application. “Analytics and data visualization, text search, graph processing, geospatial, in-memory performance and global replication allow to deliver a wide variety of real-time applications on one technology, reliably and securely” [7].

The City of Chicago uses MongoDB to create smarter and safer city. In just four months, data from Chicago’s 15 most crucial department is integrated into MongoDB which provides real-time data analysis to make better decisions. Dashboard created gives querying capability on various department data simultaneously along with twitter data for sentiment analysis [8].

MongoDB helped online retailer OTTO to improve catalog update time. OTTO now can update catalog within 15 minutes which was earlier could take 12 hours. It helps OTTO to provide a personalized experience to their customer [9].

Expedia built scratchpad app using MongoDB to personalize and help their customer by providing previous searches. It saves users all previous travel searches so that users can refer those before finalizing the travel. MongoDB’s flexible data structure and ease of development was the selling point for Expedia to use it as database store [10].

## 5 POWER OF TWO

Big data solution requires two types of technology to solve the problem, operational where real-time data is captured and stored, and analytical where this data is used for complex analysis. Frequently both of the technologies are deployed together to solve big data problem. MongoDB and Hadoop are the great choices for operational and analytical technology respectively. MongoDB can be used to store structured/unstructured data and then Hadoop MapReduce can be used to process this data for the analytics. Together they provide the complete and cost-effective solution to the big data problems [11].

## 6 CONCLUSION

Hadoop and MongoDB are the front-runner technologies to solve big data problems. The features provided by both technologies are extremely suitable to solve big data problem which requires handling of huge data and great computing power. Distributed nature of both technologies helps data to break across multiple nodes and distributed processing helps process data in parallel. Cost-effective implementation enables to accept these technologies industry-wide. There are a lot of other technologies emerging in the market but Hadoop and MongoDB will be favorite for some time to come.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Shiva Achari. 2015. *Hadoop Essentials*. Packt Publishing, Birmingham.
- [2] Apache. 2017. Apache Hadoop. web. (2017). <http://hadoop.apache.org/>
- [3] Apache. 2017. Apache Hadoop YARN. web. (2017). <http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>
- [4] Apache. 2017. HDFS Architecture. web. (2017). <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>
- [5] Apache. 2017. Powered by Apache Hadoop. web. (2017). <https://wiki.apache.org/hadoop/PoweredBy>
- [6] Guy Harrison. 2015. *Three Database Revolutions*. Apress, Berkeley, CA, Chapter 1, 3–19. [https://doi.org/10.1007/978-1-4842-1329-2\\_1](https://doi.org/10.1007/978-1-4842-1329-2_1)
- [7] MongoDB. 2017. web. (2017). <https://www.mongodb.com/mongodb-architecture>
- [8] MongoDB. 2017. web. (2017). <https://www.mongodb.com/customers/city-of-chicago>
- [9] MongoDB. 2017. web. (2017). <https://www.mongodb.com/customers/otto>
- [10] MongoDB. 2017. web. (2017). <https://www.mongodb.com/customers/expedia>
- [11] MongoDB. 2017. web. (2017). <https://www.mongodb.com/big-data-explained>

## 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

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

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

## 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 inflate your paper if you are below the page limit

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

## 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:

#### LIST OF FIGURES

1	HDFS Architecture [4]	7
2	HDFS Read [1, Ch. 3, p. 38]	8
3	HDFS Write [1, Ch. 3, p. 39]	9
4	YARN Architecture [3]	10
5	MapReduce Example [1, Ch. 3, p. 48]	11
6	Hadoop Ecosystem [1, Ch. 2, p. 26]	12
7	MongoDB Architecture [7]	13

## HDFS Architecture

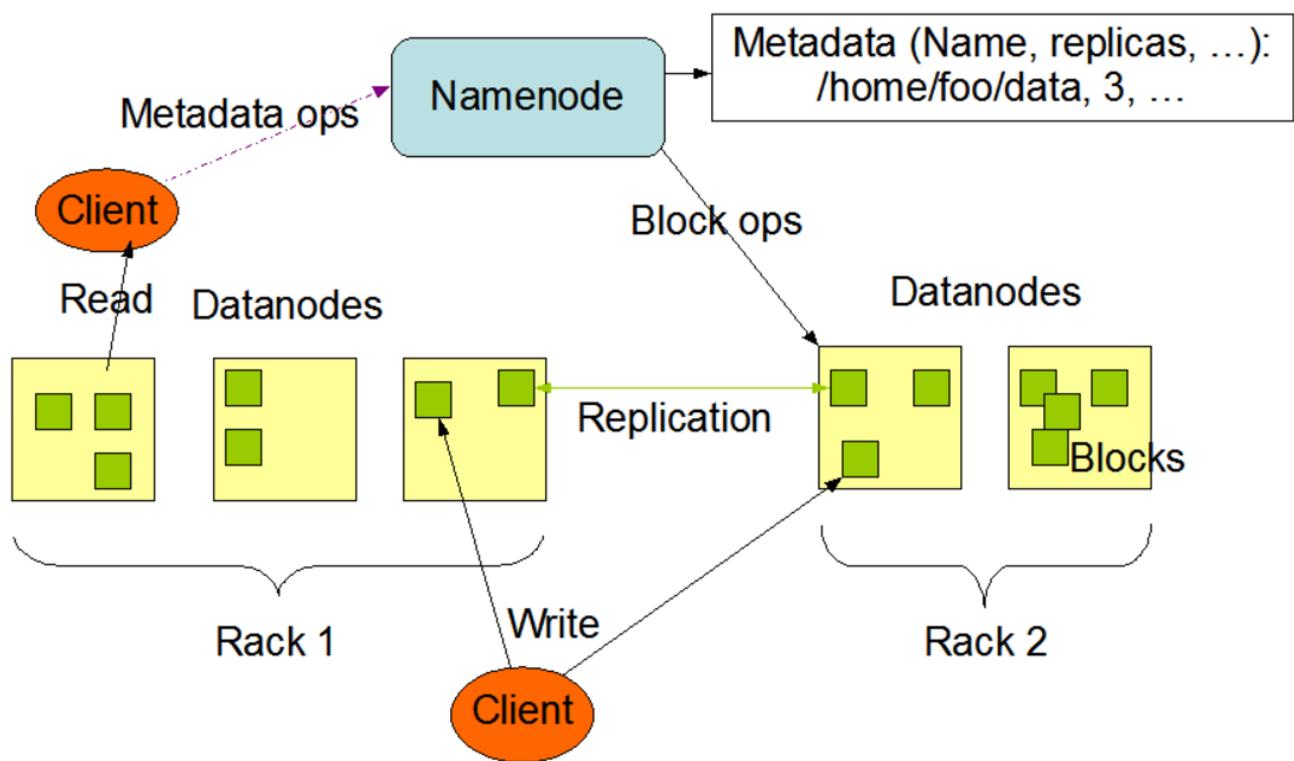


Figure 1: HDFS Architecture [4]

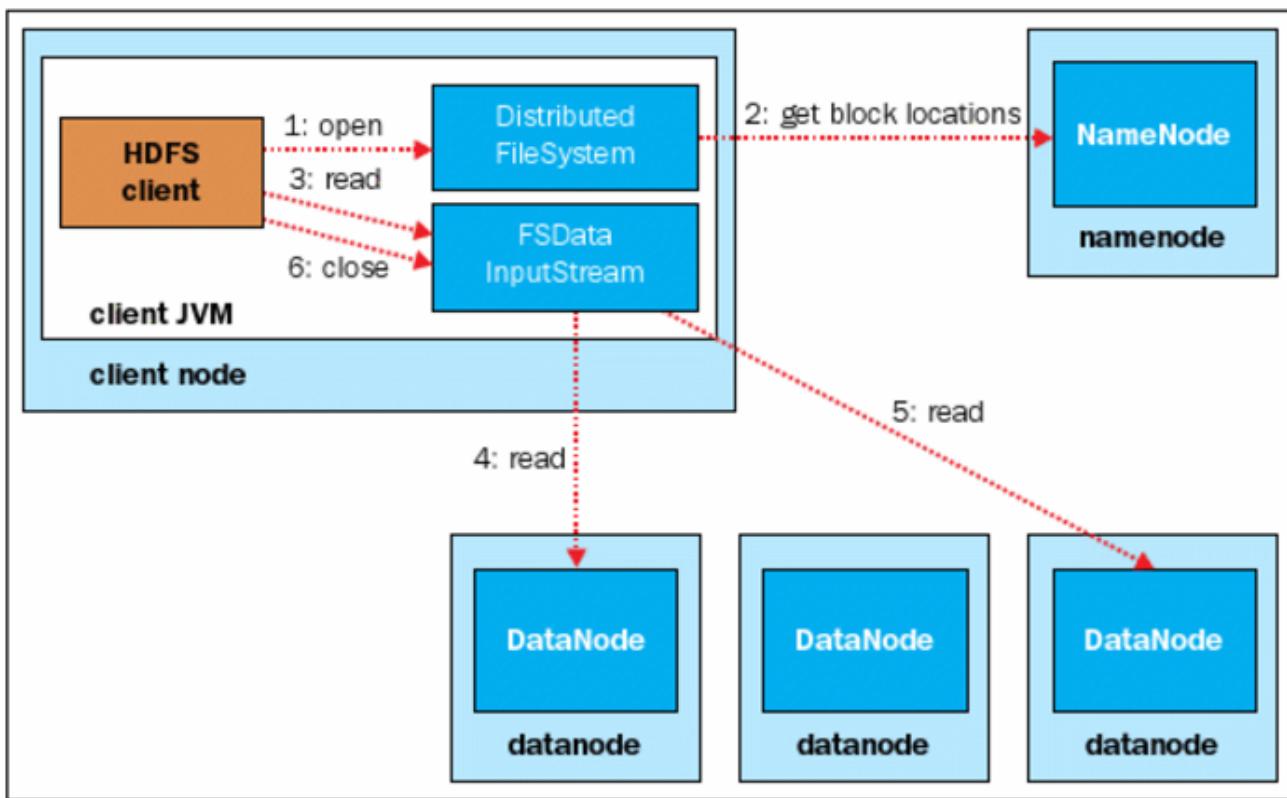


Figure 2: HDFS Read [1, Ch. 3, p. 38]

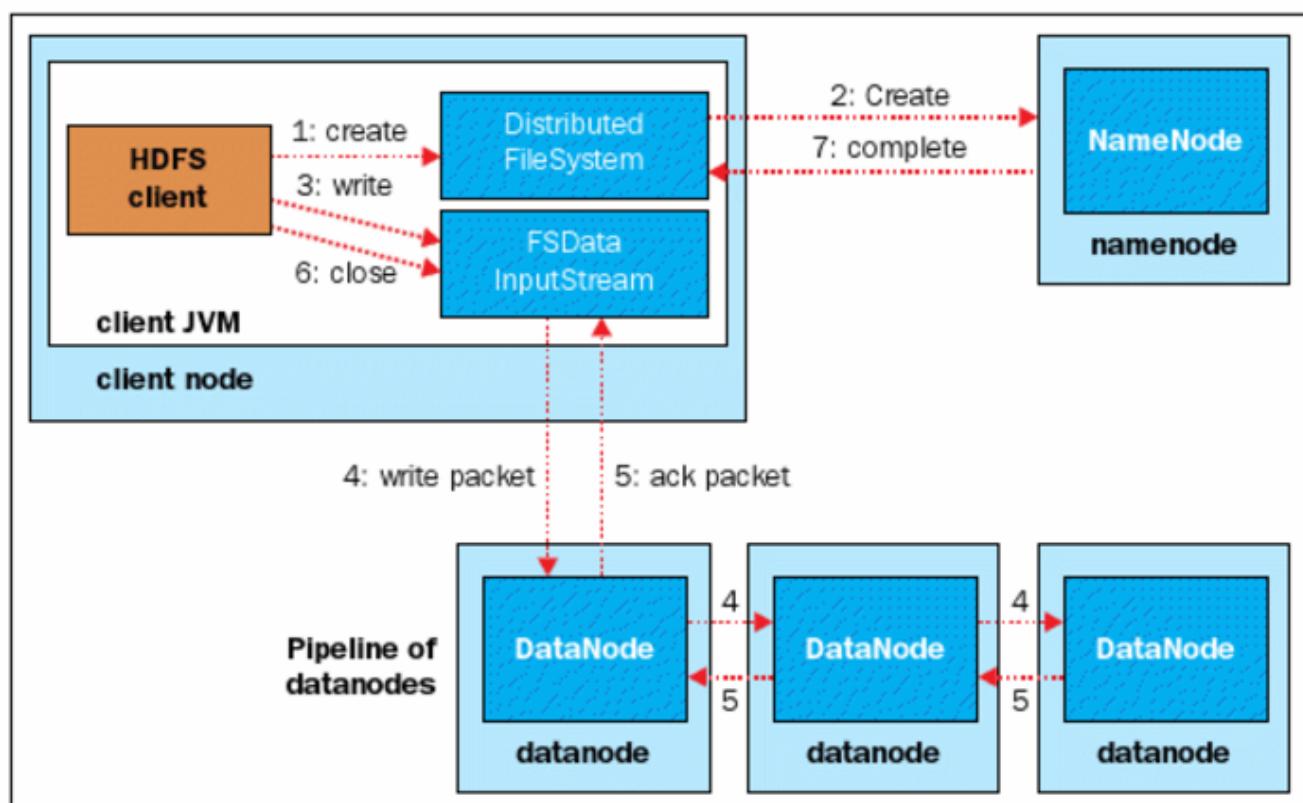


Figure 3: HDFS Write [1, Ch. 3, p. 39]

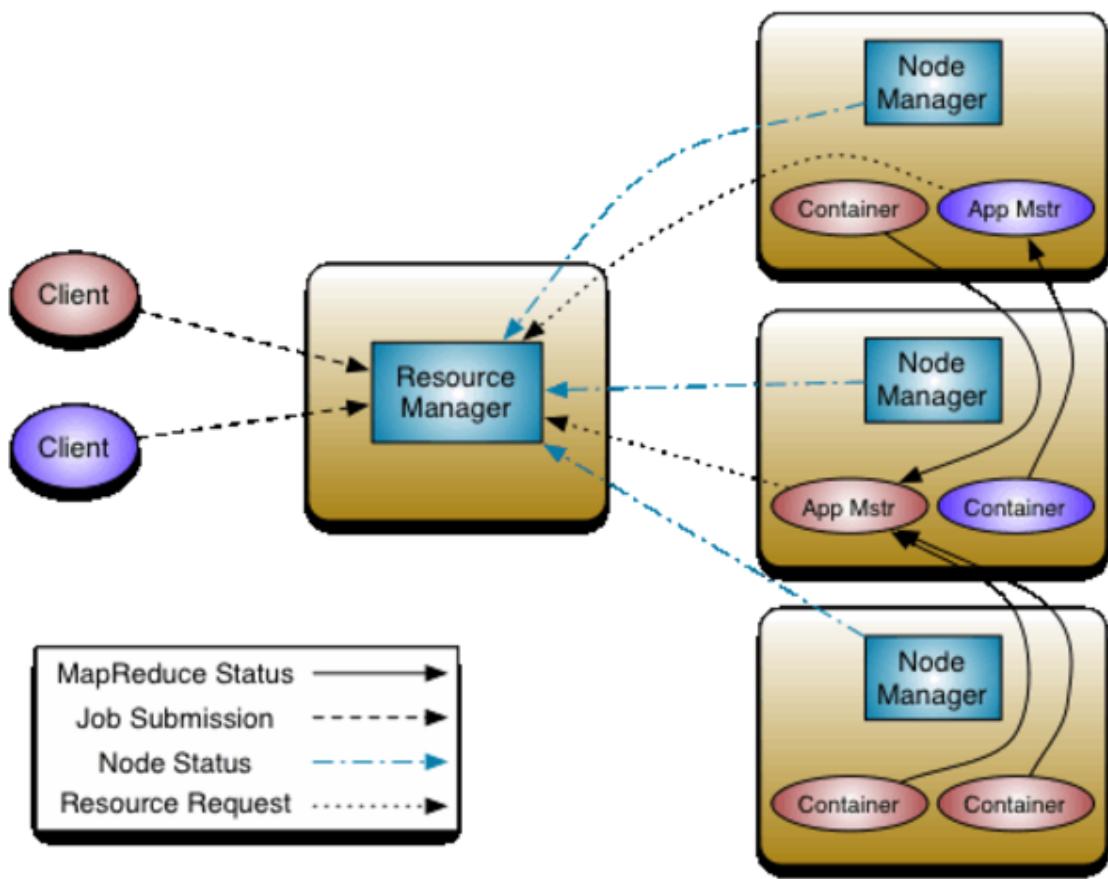


Figure 4: YARN Architecture [3]

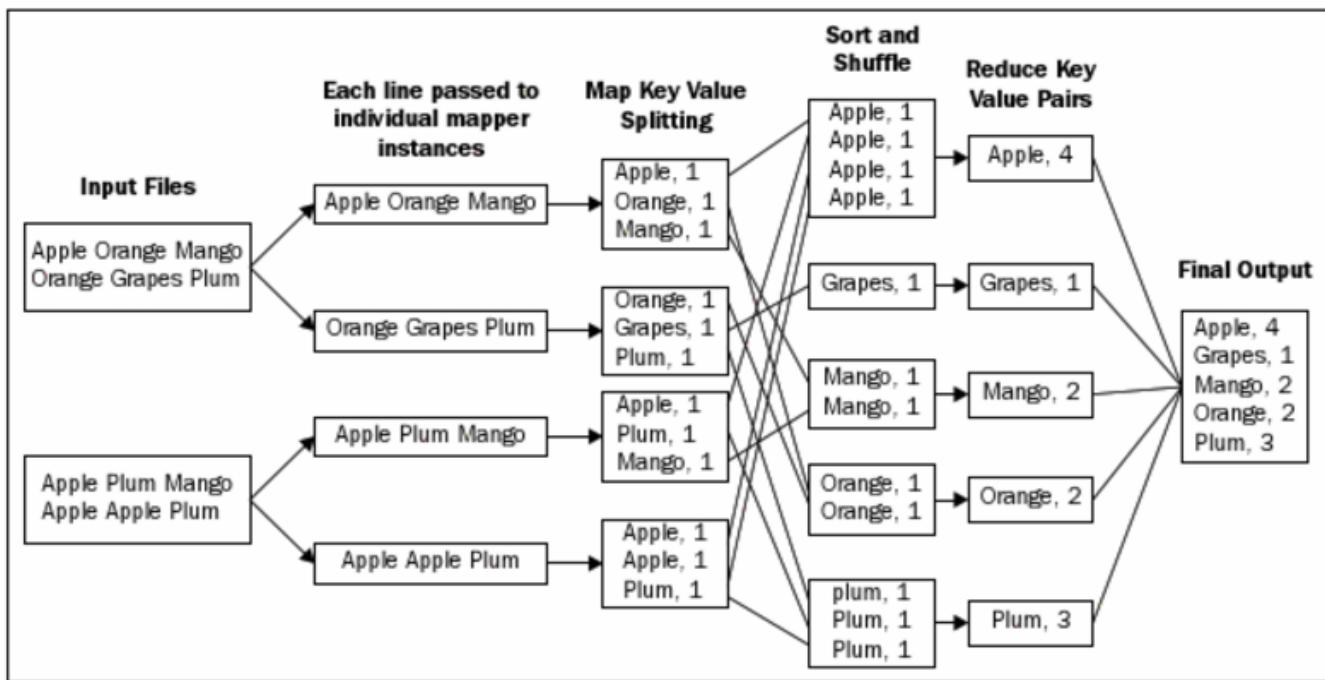


Figure 5: MapReduce Example [1, Ch. 3, p. 48]

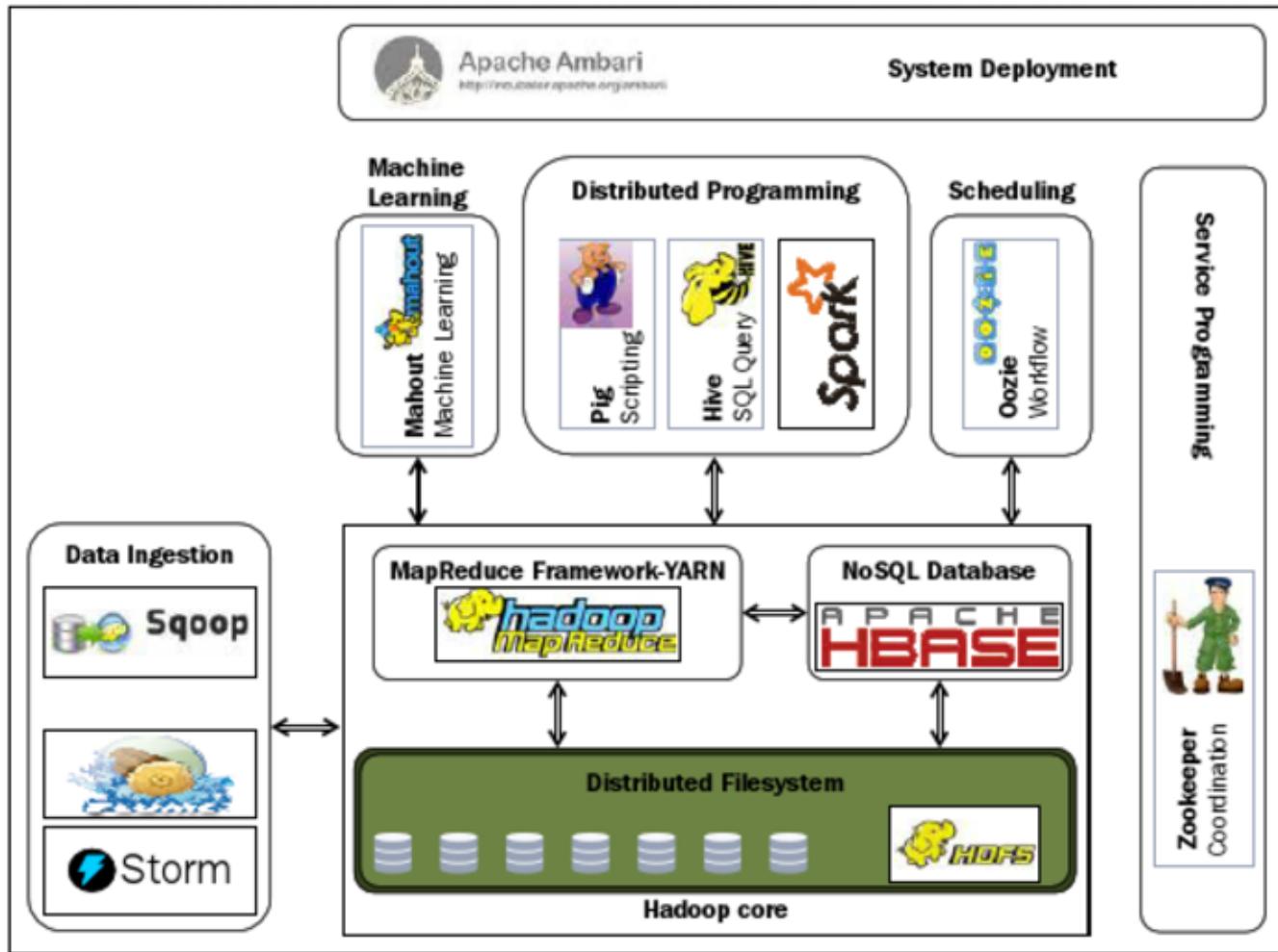


Figure 6: Hadoop Ecosystem [1, Ch. 2, p. 26]

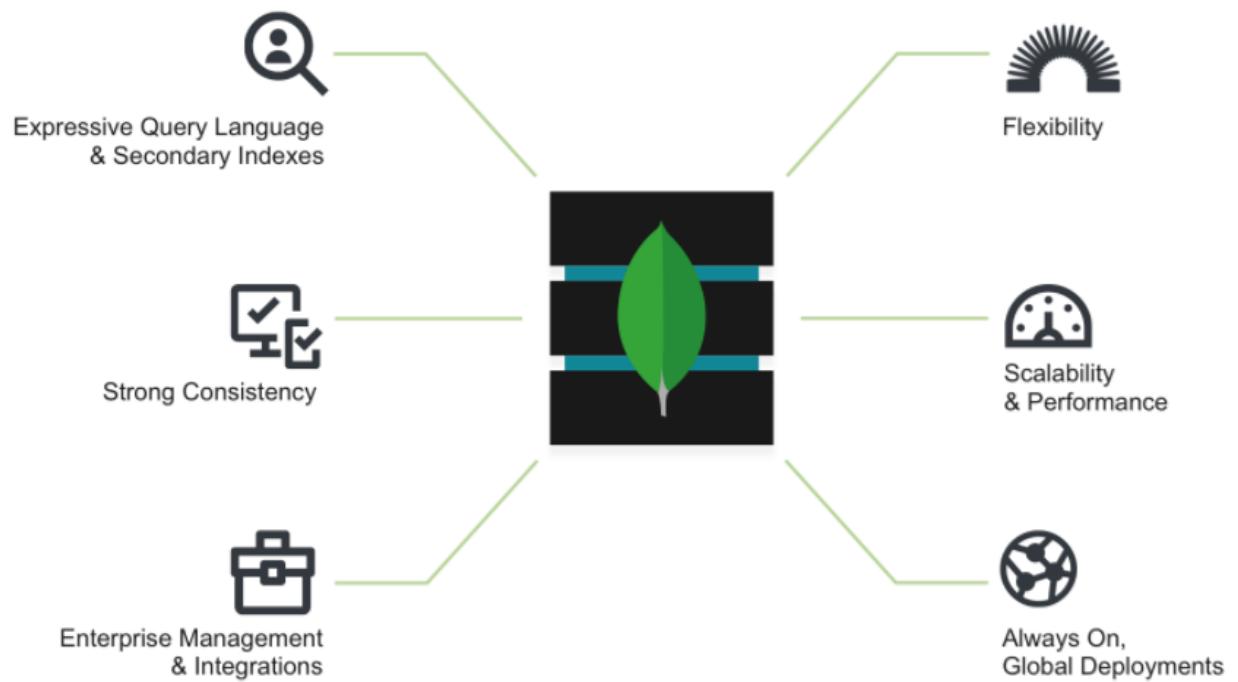


Figure 7: MongoDB Architecture [7]

## bibtex report

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

### bibtext \_ label error

bibtext space label error

bibtext comma label error

# latex report

## Compliance Report

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

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 302 paper2 (null) 2666 report.tex
wc 302 paper2 (null) 3421 report.pdf
wc 302 paper2 (null) 444 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
50: Hadoop Distributed File System (HDFS) is the default distributed
file system provided by the Hadoop. HDFS serves as storage
mechanism in the Hadoop framework. HDFC specifically designed to
process large data set and run on low-cost hardware. It is highly
fault-tolerant which contains the mechanism for quick fault
detection and auto recovery. HDFS is designed to port across
heterogeneous hardware and software platform. It does data
computation on the same node instead of moving data to the server
which is faster as well as avoid network congestion. It provides
scalability by adding or removing nodes in the HDFS cluster and
can support hundreds of nodes in single cluster \cite{www-hdfs-
arch}. Figure \ref{f:hdfs-arch} shows HDFS architecture.
```

```
51: \begin{figure}[!ht]
```

```
52: \centering\includegraphics[width=\columnwidth]{images/hdfsArch.PNG}
```

```

}
53: \caption{HDFS Architecture \cite{www-hdfs-arch}}\label{f:hdfs-arch}
56: HDFS is based on master/slave architecture where NameNode is the
   master server and DataNodes are the slave nodes. There can be only
   one NameNode server which manages file system namespace and all
   read-write requests. NameNode doesn't store any data but contains
   all the meta-data about files and DataNodes. DataNode contains
   actual data and they can be multiple in numbers usually one per
   node. DataNodes are responsible for the create, delete, replicate
   of the data blocks on the node as per the instruction by the
   NameNode. DataNode also sends block-report to NameNode which has a
   list of all blocks on the DataNode. DataNode sends the heartbeat
   message to NameNode which helps in identifying the failure nodes.
   If the heartbeat is not received by NameNode in specified interval
   then that DataNode is marked as dead and NameNode usage different
   DataNode. Figure \ref{f:hdfs-read} and \ref{f:hdfs-write} depicts
   read and write in HDFS respectively.
58: \begin{figure}[!ht]
59: \centering\includegraphics[width=\columnwidth]{images/hdfsRead.PNG}
   }
60: \caption{HDFS Read \cite[Ch.\ 3, p.
   38]{AchariShiva2015HE}}\label{f:hdfs-read}
63: \begin{figure}[!ht]
64: \centering\includegraphics[width=\columnwidth]{images/hdfsWrite.PNG}
   }
65: \caption{HDFS Write \cite[Ch.\ 3, p.
   39]{AchariShiva2015HE}}\label{f:hdfs-write}
71: Figure \ref{f:yarn-arch} shows YARN architecture.
72: \begin{figure}[!ht]
73: \centering\includegraphics[width=\columnwidth]{images/yarnArch.PNG}
   }
74: \caption{YARN Architecture \cite{www-apache-yarn}}\label{f:yarn-arch}
80: Figure \ref{f:mapreduceex} shows MapReduce process using wordcount
   example. Each line in the input file is passed to individual
   mapper class. Mapper class parses the line and sets count for the
   word. Sort and shuffle consolidate the data and sends it to the
   reducer. Reducer performs the final word count and provides the
   output.
81: \begin{figure}[!ht]
82: \centering\includegraphics[width=\columnwidth]{images/mapReduceEx.PNG}
   }
83: \caption{MapReduce Example \cite[Ch.\ 3, p.
   48]{AchariShiva2015HE}}\label{f:mapreduceex}
88: Figure \ref{f:hadoopeco} illustrates Hadoop ecosystem by various

```

```
    layers.  
89: \begin{figure}[!ht]  
90: \centering\includegraphics[width=\columnwidth]{images/hadoopEcosys  
.PNG}  
91: \caption{Hadoop Ecosystem \cite[Ch.\ 2, p.  
26]{AchariShiva2015HE}}\label{f:hadoopeco}  
105: Figure \ref{f:mongo-arch} shows MongoDB architectural  
consideration.  
106: \begin{figure}[!ht]  
107: \centering\includegraphics[width=\columnwidth]{images/mongoArch.P  
NG}  
108: \caption{MongoDB Architecture \cite{www-mongo-  
arch}}\label{f:mongo-arch}
```

```
figures 7  
tables 0  
includegraphics 7  
labels 7  
refs 6  
floats 7
```

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

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

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

---

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

label errors

bibtex errors

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

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

# Why Deep Learning matters in IoT Data Analytics?

Murali Cheruvu  
Indiana University  
3209 E 10th St  
Bloomington, Indiana 47408  
mcheruvu@iu.edu

## ABSTRACT

The Deep Learning is unique in all machine learning algorithms to analyze supervised and unsupervised datasets. Big Data challenges, such as high volumes, multi-dimensionality and feature engineering, are well addressed using Deep Learning algorithms. Deep Learning, with edge and distributed mesh computing, is best suited to handle IoT Analytics from millions of sensors producing petabytes of time-series data.

## KEYWORDS

i523, hid306, IoT, Deep Learning, Big Data Analytics

## 1 INTRODUCTION

Supervised machine learning algorithms: decision trees, linear regression, support vector machines (SVMs), Naive Bayes, neural networks, etc. are popular for classification and regression problems by analyzing labeled training data. K-means clustering algorithms are good for unsupervised datasets to categorize based on the identified patterns in unlabeled data. While there are so many factors - nature of the domain, sample size of the dataset and number of attributes defining characteristics of the data - decide which machine learning algorithm works better, Deep Learning algorithms are, getting greater traction, addressing complex analytics tasks including high-dimensionality and automatic creation of new features from existing complex hierarchical features, very well.

## 2 NEURAL NETWORKS

Neural Network is modeled after the human brain, specifically the way it solves complex problems. *Perceptron*, the first generation neural network, created a simple mathematical model or a function, mimicking neuron - the basic unit of the brain, by taking several binary inputs and produced single binary output. *Sigmoid Neuron* improved learning by giving some *weightage* to the input based on importance of the corresponding input to the output so that tiny changes in the output due to the minor adjustments in the input weights (or biases) can be measured effectively. 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 the previous layer; there will be no interaction of neurons within a layer. As shown in Figure (1), a typical Neural Network contains three layers: input (left), hidden (middle) and output (right) [3]. The middle layer is called *hidden* only because the neurons of this layer are neither the input nor the output. However, the actual processing happens in the hidden layer as the data passes through layer by layer, each neuron acts as an *activation function* to process the input. 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.

[Figure 1 about here.]

## 3 DEEP LEARNING

Deep Learning is an advanced neural network, with multiple hidden layers (thousands or even more deep), that can work well with supervised (labeled) and unsupervised (unlabeled) datasets. Applications, such as speech, image and behavior patterns, having complex relationships in large-set of attributes, are best suited for Deep Learning Neural Networks. Deep Learning vectorizes the input and converts it into output vector space by decomposing complex geometric and polynomial equations into a series of simple transformations. These transformations go through neuron activation functions at each layer parameterized by input weights. For it to be effective, the cost function of the neural network must guarantee two mathematical properties: *continuity* and *differentiability*.

[Figure 2 about here.]

### 3.1 Feature Engineering

The dataset with too many dimensions, also known as attributes or features, create large sparsity and make it difficult to process. *Curse of dimensionality* is a scenario where the value added by the dimensions is much smaller in comparison to the processing cost. However, in certain applications, such as face recognition and patient electronic medical records, the complexity created by multiple dimensions might add value to the context. *Feature Engineering* is an exploratory analysis to identify the features that collectively contribute to better predictive modeling by removing irrelevant features and creating new features, using the training information to identify the patterns, from existing interrelated features [6]. *Principal component analysis* (PCA) is a technique to analyze the interdependency among the features and keep only the principal, most relevant, features with minimum loss in the model. With enough training, Deep Learning makes neurons learn new features themselves, in an unsupervised manner, from existing features distributed in several hidden layers. *Stacked Autoencoder* (AE) is, a Deep Belief Network algorithm, to create advanced predictive models for large datasets having thousands or even millions

of dimensions, automatically, with complex hierarchical attributes in non-linear fashion for simpler computing. Though AE is sophisticated, it is very difficult to understand the algorithm logic and so unable to reuse the learnings from the modeling to other systems.

### 3.2 Deep Neural Networks

*Convolutional Neural Network* (CNN), also called multilayer perceptron (MLP), is a deep feedforward network, consists of (1) convolutional layers - to identify the features using weights and biases, followed by (2) fully connected layers - where each neuron is connected from all the neurons of previous layers - to provide non-linearity, sub-sampling or max-pooling, performance and control data overfitting [2]. CNN is used in image and voice recognition applications by effectively using multiples copies of same neuron and reusing group of neurons in several places to make them *modular*. CNNs are constrained by *fixed-size* vectorized inputs and outputs. *Recursive Neural Network* (RNN) is, another type of Deep Learning, that uses same shared feature weights recursively for processing sequential data, emitted by sensors or the way spoken words are processed in natural language processing (NLP), to produce arbitrary size input and output vectors. RNN uses a technique called *loop*, where several copies of the same chunk of network (module), each instance passing a message to the next, to persist the information. Long Short Term Memory (LSTM) is an advanced RNN to learn and remember *longer* sequences by composing series of repeated modules of neural network and a concept called *cell state*, a memory unit, to memorize the learning by adding and removing information using *input*, *output* and *forget* gates, in a regularized fashion while data flows through the layers [9]. The Convolutional and Recursive Neural Networks can complement each other to produce better and effective models where problem space has both - hierarchical features and temporal data. Deep Learning can also work well with related *Reinforcement Learning* algorithms where the focus is on how to maximize the learning based on rewards and punishments.

[Figure 3 about here.]

[Figure 4 about here.]

## 4 IOT DATA ANALYTICS

Internet of Things (IoT) is getting lots of traction, due to the massive volumes and variety of the sensor data, qualifying it to be part of *Big Data*; however, business needs to convert this data into *information* whether to monitor and control the things (devices) or to analyze the sensor data for betterment. Time-series data has non-stationary time aspects collected at certain intervals over a short period of time and correlate this sequence of data with past or future sequences. Stock prices and IoT sensor data are examples of time-series data. *InfluxDB*, an open source time-series database, is offering high write performance, data compaction through down-sampling and automatic deletion of expired old time-series data, to address IoT data storage challenges [5].

### 4.1 Complexity

Unique traits of IoT data, such as noise, high dimensionality and high streaming of time-series data in real-time, make it challenging

to process using traditional machine learning algorithms [10]. Autoregressive Moving Average Model (ARIMA), converts time-series from non-stationary into stationary, but only for short-time predictions. Deep Learning, using LSTM, can detect anomalies in the sensor data and train time-series patterns very well. Deep Learning algorithms involve complex mathematics - geometry, matrix algebra, differential calculus, statistics and probability, and intensive distributed computing to train the massive amounts of sensor data.

### 4.2 Scalability

Deep Learning, by design, allows parallel programming, as each module - with all the dependencies among neurons - can run independently and parallelly from other modules within the network. Using Graphics Process Unit (GPU), module networks can achieve parallel programming without needing much of Central Processing Unit (CPU) allocation of a computer. Though GPU is intended for graphical processing, it works efficiently to run thousands of small mathematical functions, such as matrix multiplications, in parallel. Cloud computing and edge analytics offer flexible scale out distributed processing options using virtualization and containerization. Sophisticated algorithms and distributed computing make Deep Learning scale and perform well to process huge datasets.

### 4.3 Case Study

Hewlett Packard (HP) Labs has given a presentation of their research to measure the effectiveness Deep Learning algorithms on IoT Sensor Data Analytics. Sample data - vision, speech, text and sensor signals, has been collected from scripted video and the accelerometer from 52 subjects gathered 20 minutes of activity recognition per subject averaging 12,000 measurements per minute per person with 16 classifications, such as walk to bed, enter bed, lie down, roll left, roll right and speak. They have analyzed and trained the sample time-series data using various supervised learning algorithms including SVMs, decision trees and traditional neural networks; compared the results with recurrent, Deep Learning, neural network. Deep Learning showed 95% or more accuracy in various scenarios, performed much better than all the other algorithms, without sophisticated feature engineering. However, Deep Learning algorithms were predictively slow and expensive for results to converge as the sample dataset is huge with lots of instances ( $10^6$ - $10^9$ ) and very large number of features ( $>10^6$ ). They have concluded the presentation with scale-out hardware options using CPU/GPU clusters, edge analytics and futuristic distributed mesh computing alternatives for better scalability and performance [11].

## 5 CONCLUSION

In contrast to traditional machine learning solutions, Deep Learning not only scales well with high volumes of input data but also facilitates in automatic decomposition of complex data representations of unsupervised and uncategorized data. Automatic discovery of new features, from convolutional or recurrent neural networks, makes Deep Learning predominant among all machine learning algorithms. It is very difficult to understand fuzzy and complex logic of Deep Learning; perhaps, more adoption helps getting better handle at them. Deep Learning algorithms need deep research in

validating the process of advanced Big Data Analytics tasks, such as IoT sensor time-series data, semantic learning, scalability, data tagging and reliability of the predictive models without extreme generalization.

## ACKNOWLEDGMENTS

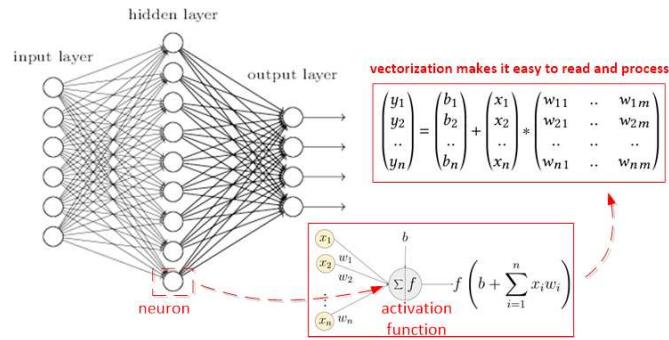
The author would like to thank Dr. Gregor von Laszewski and the Teaching Assistants for their support and valuable suggestions.

## REFERENCES

- [1] Mark Chang. 2016. Applied Deep Learning 11/03 Convolutional Neural Networks. (Oct. 2016). <https://www.slideshare.net/ckmarkohchang/applied-deep-learning-1103-convolutional-neural-networks>
- [2] Christopher Olah. 2014. Conv Nets: A Modular Perspective. (July 2014). <http://colah.github.io/posts/2014-07-Conv-Nets-Modular/>
- [3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>
- [4] Vikas Gupta. 2017. Understanding Feedforward Neural Networks. (Oct. 2017). <https://www.learnopencv.com/understanding-feedforward-neural-networks/>
- [5] Influx. [n. d.]. *InfluxDB is the Time Series Database in the TICK stack*. Technical Report. Influx. <https://www.influxdata.com/time-series-platform/influxdb/>
- [6] Jason Brownlee. 2014. Discover Feature Engineering, How to Engineer Features and How to Get Good at It. (Sept. 2014). <https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>
- [7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. (May 2015). [http://www.nature.com/nature/journal/v521/n7553/fig\\_tab/nature14539\\_F5.html](http://www.nature.com/nature/journal/v521/n7553/fig_tab/nature14539_F5.html)
- [8] Nicholas Leonard. 2016. Language modeling a billion words. (July 2016). <http://torch.ch/blog/2016/07/25/nce.html>
- [9] Christopher Olah. 2015. Understanding LSTM Networks. (Aug. 2015). <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [10] Rajesh Sampathkumar. 2016. Time Series Analysis of Sensor Data. (Aug. 2016). <http://www.thedatateam.in/time-series-analysis-of-sensor-data/>
- [11] Natalia Vassilieva. 2016. *Sense Making in an IOT World: Sensor Data Analysis with Deep Learning*. Technical Report. Hewlett Packard Labs. <http://on-demand.gputechconf.com/gtc/2016/presentation/s6773-natalia-vassilieva-sensor-data-analysis.pdf>

LIST OF FIGURES

1	Simple Neural Network [3, 4].	5
2	Deep Neural Network with three hidden layers [3].	5
3	Sample Convolutional Neural Network [1].	5
4	Recursive Neural Network Loop and LSTM Cell State [7, 8].	6



An example of a neuron showing the input ( $x_1 - x_n$ ), their corresponding weights ( $w_1 - w_n$ ), a bias ( $b$ ) and the activation function  $f$  applied to the weighted sum of the inputs.

Figure 1: Simple Neural Network [3, 4].

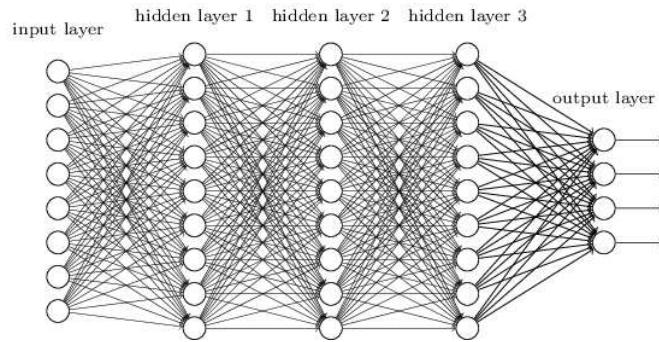


Figure 2: Deep Neural Network with three hidden layers [3].

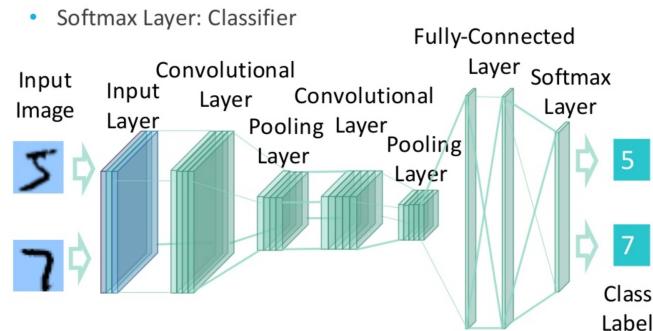
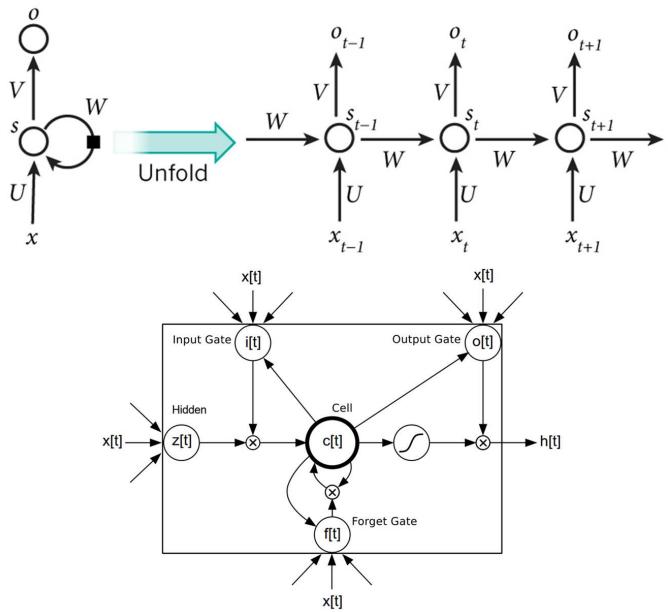


Figure 3: Sample Convolutional Neural Network [1].



**Figure 4: Recursive Neural Network Loop and LSTM Cell State [7, 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
Warning--empty address in Goodfellow2016
Warning--empty year in Influx
(There were 2 warnings)
```

```
bibtext _ label error
```

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

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

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

```
[2017-11-06 17.37.11] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.0s.
./README.yml
8:63      error    trailing spaces  (trailing-spaces)
9:63      error    trailing spaces  (trailing-spaces)
11:61     error    trailing spaces  (trailing-spaces)
12:60     error    trailing spaces  (trailing-spaces)
26:67     error    trailing spaces  (trailing-spaces)
27:66     error    trailing spaces  (trailing-spaces)
28:63     error    trailing spaces  (trailing-spaces)
29:53     error    trailing spaces  (trailing-spaces)
30:62     error    trailing spaces  (trailing-spaces)
31:61     error    trailing spaces  (trailing-spaces)
32:55     error    trailing spaces  (trailing-spaces)
42:10     error    too many spaces after colon  (colons)
```

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

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

```
yamlcheck
-----
```

```
wordcount
-----
```

```
6
wc 306 paper2 6 1849 report.tex
wc 306 paper2 6 1931 report.pdf
wc 306 paper2 6 273 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
-----
```

```
passed: True
```

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

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

```
passed: True
```

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

```
45: \begin{figure}
47: \includegraphics[width=0.5\textwidth]{images/neuralnetwork}
48: \caption{Simple Neural Network \cite{Goodfellow2016, Gupta2017}.}
```

```
\label{fig:figure1}
56: \begin{figure}
58: \includegraphics[width=0.5\textwidth]{images/deepnetwork}
59: \caption{Deep Neural Network with three hidden layers
   \cite{Goodfellow2016}.} \label{fig:figure2}
71: \begin{figure}
73: \includegraphics[width=0.5\textwidth]{images/cnn}
74: \caption{Sample Convolutional Neural Network \cite{Chang2016}.}
   \label{fig:figure3}
77: \begin{figure}
79: \includegraphics[width=0.5\textwidth]{images/rnn}
80: \caption{Recursive Neural Network Loop and LSTM Cell State
   \cite{LeCun2015, Leonard2016}.} \label{fig:figure4}
```

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

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

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

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

---

```
find textwidth
```

```
47: \includegraphics[width=0.5\textwidth]{images/neuralnetwork}

58: \includegraphics[width=0.5\textwidth]{images/deepnetwork}

73: \includegraphics[width=0.5\textwidth]{images/cnn}

79: \includegraphics[width=0.5\textwidth]{images/rnn}

passed: False
```

below\_check

---

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty address in Goodfellow2016
Warning--empty year in Influx
(There were 2 warnings)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

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

find newline

---

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

```
find cite {
```

---

```
-----
```

```
passed: True
```

# Big Data in Decentralized election

Ashok Kuppuraj

Indiana University

Bloomington, Indiana 43017-6221

akuppura@iu.edu

## ABSTRACT

In the current world of technology, lots of legacy practices are modernized, some are yet to modernize and some are extinct. Though lots of inventions and modernization are happening in every spectrum of life, the election process in a democratic political system is yet to attain the quantum of modernization. In the era of Big data and technologies, how the election process can be made efficient, accountable and decentralized.

## KEYWORDS

i523, hid324, Big data, Election, India, U.S, Blockchain, Infrastructure

## 1 INTRODUCTION

The election is a formal decision-making process by a population to make a representation of them in a democratic system, this process of electing an individual is called as Representative Democracy. The election is one of the important activity in fulfillment of democracy, it is based on the fact that "Majority rule"[8], the theory holds good from cells in the human body to the transaction validations in Blockchain. In the current digital world, the election is still a slow, untrustworthy process. With all the implementation aspects, Big data and Blockchain can help modernize the election process to more secure, trustworthy, foolproof, instant and decentralized process. In today's world, the election results are impacted by big data why can't the election itself impacted by it for its own good.

## 2 ELECTION IN INDIA AND U.S

The election is the systematic process of selecting an individual to represent an entire population, though some countries don't follow the same process, most of the countries have a process of election[15]. A good example to consider for democratic countries is India and U.S.A, the former one is the largest democracy in the world and the later one is the oldest democracy in the world. Especially in India, with a population of more than 1 billion[2], execution of election is a tedious and costly process. In the beginning of this decade, both the countries witnessed the implications of Big data and its technologies along with Internet made a decisive role in the outcome of the election. In the year 2014 and 2016 for India and U.S respectively, political parties won the election with the help of Big data and analytics[10].

### 2.1 Big data in Indian Election

India is known for its diversity in terms of population, language, and culture. Conducting election to such a diversified country itself is a big challenge with the current technological advancement. For an instance lets consider the size of the Indian electorate, with the sheer volume of 814.5 million voters from 29 states and 12 different

languages is a good use case for Big data[11]. In a political party perspective, they had to process millions of Information sets from Twitter, Facebook to browser cookies and newspaper sales data to understand the right place for canvassing, raising funds and improving the face value, etc. And all these have to be done at a specific point of time to derive a relevant output. In an Election commission's perspective, which is an independent authority to conduct the election, the data generation starts from the first day of announcing the election date. It begins with applications of the contestants in multiple languages, its validation, voters IDs, EVM(Electronic Voting machine) data, votes aggregation, validation all involves a tremendous amount of data.

### 2.2 Big data in U.S Election

U.S is one of the advanced countries in terms of the election process. Similar to any other democratic country, it has Federal election commission which will conduct elections, with 115 million voters and 87.36% internet penetration [9], data generated for the campaigning, voter validation, polling adds up to the big data use case. The aftermath of 2016 US election showed the prowess of Big data and analytics. For the predictions and campaigns, political parties amassed more than 5000 data points about the behavioral patterns from Healthcare to car ownership data, purchased Digital trails, Facebook behavioral patterns to target potential voters, increase the awareness penetration and predict the results with data points backing it[13].

In both countries, the digital imprint of elections was around big data and its technologies. However, the technology has only reached a single side of the river. The Political campaign, advertisement started using all sorts of advanced technology, but the voting itself hasn't been improved, it is still slow to implement, results take several days to announce and very costly to implement over a vast region.

## 3 PROBLEMS WITH ELECTION PROCESS

Before giving a solution in Big data, what are the possible problems in election, who are the stakeholders impacted and what is the integrity of an election result, all these come into the picture. Here, we group the problems by stakeholders. There are three stakeholders in the election process, first is the Independent agency conducting the election, Political party and the people, who cast their vote.

In the people's perspective, the common problems are, they have to travel to a common place in their locality and stand in a long queue to cast vote and might be deceived by advertisement, biased media and they select their representative purely based on trust. Once the vote is cast, there is no way for the people to revert it or alter it and the politician don't have any liability till the next term. In practical terms, there are no means for the people to evaluate a

candidate post-election performance and the frequency to do the same is so high that the people have to wait for the next term, which will be 4-5 years.

In Election commission perspective, their sole purpose is to maintain the integrity of the election, so that the majority's decision reflects in the result. The major problem is the authenticity of voters and contestants, second is logistics and communication,i.e bringing the people from geographically and culturally location on common terms, third is to make sure the casted votes are untampered and the last one delayed delay in result announcement.

In contestant perspective, all contestant should have a level playing ground irrespective of the competition's fame and wealth.

And, the turnout volume of election is low that there is a high possibility that its base motivation might fail. If the turnout is less than 50%, then there is a high possibility that the entire election might go wrong. For example, in 2016 U.S election, the turnout is 55.5% [7], in India, the same in 2014 is 66.40 % [5].

## 4 BIG DATA IN DECENTRALIZED ELECTION

As election is a staged approach, its solution would be staged as well.

### 4.1 Voter/Contestant selection

In the Big data terms, the voter selection can be synonyms with data ingestion into a data lake or no-SQL database. Technically, we have to persist not more than 1.2 billion records, considering the population of India. By efficiently sharding it per state, we can easily persist such volume of data into a distributed storage. This is already implemented in some of the countries like U.S in the name of SSN [12] and in India, it is Aadhaar ID[1]. With the availability of all data, we can easily read and filter out the voters based on their criminal records whether they are eligible for voting or contesting or not.

### 4.2 Voting

Voting, in simple terms, can be associated with aggregation/summing based on the key. Here the key is the contestant's symbol or name. By hosting this voting process in a website portal with API calls and load balancers, we can stream the votes and aggregate while it streamed to a persistent data store. With this approach, we can decentralize voting process. We can deter abusing this model by windowing the voting time. Upon completion of the window, we can reuse the infrastructure for other e-governance projects or we can reuse the e-governance infrastructure for this by using YARN or other third party tools as the resource manager, this can be possible even with streaming apps. Per the benchmarking done at MongoDB, we can attain up to 100k/second inserts [3]. By optimizing the insert, load balancing the API and sharding based on different mediums and methods, we can build an architecture to withstand such high load in a short span of time. However, the validation of voters has to be completed before casting.

### 4.3 Election result

In continuation with the voting implementation of big data technologies, results announcement is just an aggregation call over the database. In the current world, the vote calculation takes a day to

announce the results. With the big data in place, the result can be published on the same day or in near real time.

### 4.4 Election Frequency

Election frequency is proportional to the terms of the contestant for a given position. As the democratic principle believes that the people rule themselves, what if the representative after winning the election did not fulfill the expectations. The people have to wait for the next term to make any change and it is not feasible financial wise to do it immediately. With the big data in place and resource being available, we can increase the frequency of election, so that the representative is accountable for the promises. For example, if we have a terms set for 5 years, every year once, performance verification can be done in the form of a negative vote to the selected contestant and if the count is less than 50% of his/her total vote count, the contestant can be disqualified.

## 5 BLOCKCHAIN IN DECENTRALIZED ELECTION

As the Blockchain is known for the security and reliability, it can be leveraged along with big data technologies to implement a secure election on a decentralized infrastructure. The idea is that all the eligible voter will be provided with a token before a day of actual polling. When the window for polling begins, you can transfer the token with the candidate's value, it can be a number or a code to a common address. Upon calculation of valid ones, the token can be sent back for the next set of the election. As the blockchain is protected mathematically, we can ensure the authenticity of voting and an individual can be sure that his/her vote is a validated one. Also, an audit trail can be persisted to check on voting fraud. The election commission, can easily validate the tokens, aggregate the votes and announce the results[6]. For example, FollowMyVote proposes voting entirely on Blockchain. The anonymity of voter is maintained by Elliptic curve Cryptography[14], the transaction, and consensus similar to Bitcoin network. And, BitCongress propose a system combining Bitcoin, Counterparty and Smart contracts. It proposes a token called VOTE, which can be transferred to the contestants and by the end of the election, the token will be transferred back [4].

Though the stability of Blockchain over the volume of a national election is not well tested or implemented. By augmenting with the Big data technologies like streaming and in-memory processing, this can be achieved in future. Once established, multiple countries can conduct an election on a single infrastructure without the fear of hacking or tampering.

## 6 CONCLUSION

Though the election process is evolving in a pace different from the current world, there is a desperate need to modernize it to continue its legacy of giving people their right. To synchronize it with the current advancement in other areas, Big data and Blockchain can be leveraged. Maybe in the future, we do not need representatives for us, instead, our collective decisions may be taken forward as actual decision with Artificial Intelligence.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Aadhaar. 2014. UIDAI - Official Website. <https://uidai.gov.in/main-content>. (2014). (Accessed on 11/05/2017).
- [2] Central Intelligence Agency. 2017. The World Factbook fl? Central Intelligence Agency. [https://www.cia.gov/library/publications/the-world-factbook/docs/contributor\\_copyright.html](https://www.cia.gov/library/publications/the-world-factbook/docs/contributor_copyright.html). (2017).
- [3] Sam Bhat. 2015. *High Performance Benchmarking: MongoDB and NoSQL Systems – MongoDB*. Technical Report 2. United Software Associates Inc, 5674 Stoneridge Dr 100, Pleasanton, CA 94588.
- [4] BitCongress. 2016. BitCongressWhitepaper.pdf. <https://bravenewcoin.com/assets/Whitepapers/BitCongressWhitepaper.pdf>. (2016). (Accessed on 11/06/2017).
- [5] Election commission of India. 2014. RptPC\_WISE\_TURNOUT. [https://web.archive.org/web/20140606003137/http://eci.nic.in/eci\\_main1/GE2014/PC\\_WISE.TURNOUT.htm](https://web.archive.org/web/20140606003137/http://eci.nic.in/eci_main1/GE2014/PC_WISE.TURNOUT.htm). (2014). (Accessed on 11/05/2017).
- [6] Ming Chow Francesca Caiazzo. 2016. *A Block-Chain Implemented Voting System*. Technical Report. Tufts University. 6–9 pages.
- [7] Michael P. McDonald. 2016. Associate Professor, University of Florida. <http://www.electproject.org/2016g>. (2016).
- [8] Anthony J. McGann. 2002. The Tyranny of the Super-Majority: How Majority Rule Protects Minorities - eScholarship. <https://escholarship.org/uc/item/18b448r6author>. (2002).
- [9] United Nation. 2014. UNdata – record view – Percentage of individuals using the Internet. <http://data.un.org/Data.aspx>. (2014). (Accessed on 11/05/2017).
- [10] NBC News. 2016. How Big Data Broke American Politics - NBC News. <https://www.nbcnews.com/politics/elections/how-big-data-broke-american-politics-n732901>. (2016). (Accessed on 11/05/2017).
- [11] Furhaad Shah. 2014. The First Prime Minister to Use Big Data. <http://dataeconomy.com/2014/05/narendra-modi-first-prime-minister-use-big-data-analytics/>. (May 23 2014). (Accessed on 11/05/2017).
- [12] SSN. 2017. Social Security Number and Card – Social Security Administration. <https://www.ssa.gov/ssnumber/>. (2017). (Accessed on 11/05/2017).
- [13] Gillian Tett. 2016. Trump, Cambridge Analytica and how big data is reshaping politics. <https://www.ft.com/content/e66232e4-a30e-11e7-9e4f-7f5e6a7c98a2>. (2016). (Accessed on 11/03/2017).
- [14] Follow My Vote. 2016. Elliptic Curve Cryptography In Online Voting - Follow My Vote. <https://followmyvote.com/online-voting-technology/elliptic-curve-cryptography/>. (2016). (Accessed on 11/05/2017).
- [15] Wikipedia. 2016. Elections by country - Wikipedia. [https://en.wikipedia.org/wiki/Elections\\_by\\_country](https://en.wikipedia.org/wiki/Elections_by_country). (2016). (Accessed on 11/05/2017).

## 7 ISSUES

DONE:

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

### 7.1 Assignment Submission Issues

DONE:

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

### 7.2 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

### 7.3 Formatting

DONE:

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

DONE:

Other formatting issues

### 7.4 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}

### 7.5 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

## 7.6 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

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
```

## 7.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

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

## 7.8 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:

```
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
```

```
[2017-11-06 17.37.36] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Typesetting of "report.tex" completed in 1.3s.
./README.yml
 8:81      error    line too long (142 > 80 characters) (line-length)
 20:81     error    line too long (117 > 80 characters) (line-length)
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Ashok Kuppuraj
hid: 324
```

```
paper1: 100% Oct 31 17  
paper2: 100% Nov 6 17
```

```
yamlcheck
```

---

```
wordcount
```

---

```
4  
wc 324 paper2 4 1948 report.tex  
wc 324 paper2 4 2801 report.pdf  
wc 324 paper2 4 455 report.bib
```

```
find "
```

---

```
39: The election is one of the important activity in fulfillment of  
democracy, it is based on the fact that "Majority  
rule"\cite{1:online}, the theory holds good from cells in the  
human body to the transaction validations in Blockchain. In the  
current digital world, the election is still a slow, untrustworthy  
process. With all the implementation aspects, Big data and  
Blockchain can help modernize the election process to more secure,  
trustworthy, foolproof, instant and decentralized process. In  
today's world, the election results are impacted by big data why  
can't the election itself impacted by it for its own good.
```

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

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

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

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

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Analytics and Insurance Fraud Detection

Qiaoyi Liu

Indiana University of Bloomington

3209 E 10th St

Bloomington, Indiana 47408

ql30@umail.iu.edu

## ABSTRACT

This paper is to analysis how people using big data to detect Insurance Fraud in real life.

## KEYWORDS

i423, hid106, Data Science, Big Data Analytics, Cloud Computing, fraud detection

## 1 INTRODUCTION

Digitization set apart by an increasing number social media and mobile devices is shifting the business landscape in every sector insurance included. The opportunity presented by this aspect for insurance companies are immense. Communities and social networks enable insurers to interface with their clients better, which to their advantage improves branding, customer retention, and acquisition [? ]. Insurance companies additionally get a plenty of contributions from computerized data as feedbacks, which likewise can be utilized to develop unique products and aggressive valuing. Digitization of big data analytics offers numerous opportunities that Insurances Company can harness to detect fraud among their customers. Dealing with fraud manually has dependably been expensive for insurance firms regardless of the possibility that maybe a couple of minor fraud went undetected [? ]. What's more, the trends in big data (the evolution in unstructured information) are prone to numerous fraud, which can go without notice if analysis is performed correctly. In the proceeding section, the article will examine important of big data in insurance fraud detection and its relevancy.

## 2 IMPORTANCE BIG DATA AND INSURANCE FRAUD DETECTION

Conventionally, insurance firms utilize statistical models to recognize fraudulent cases. These models have their limitation [? ]. To start with, they employ sampling techniques to assess information, which prompts at least one fraud going unnoticed. There is a punishment for not performing a proper assessment of the data provided. Subsequently, this strategy depends on the cases analyzed before. Therefore, every time different fraud takes place, insurance firms need to manage the impact for the first time. Lastly, the conventional strategy works in silos and is not correctly equipped for taking care of the natural developing wellsprings of data from various diverts and diverse capacities in an integrated way. Analytics tends to be difficult and assumes an exceptionally pivotal part in fraudulent recognition for insurance firms. In the proceeding section, the significant benefits of utilizing big analytics in fraud detection assessed.

### 2.1 Identification of low incidence events:

Utilizing sampling methods accompanies its particular arrangement of acknowledged mistakes. By using analytics, insurance can manufacture frameworks that go through every fundamental datum. This like this distinguishes events with low frequency (0.001%) [? ]. Methods such as predictive modeling can be utilized to altogether break down processes of fraud, channel clear cases, and allude low-rate fraud cases for facilitating analytics.

### 2.2 Enterprise-wide solution:

Analytics help in building a global point of view of the anti-fraud endeavors all through the undertaking. Such a point of view regularly prompts dominant fraud location by connecting related data inside the association. Fraud can happen at various source focuses premium, claims or surrender, application, employee-related or outsider fraud. In the meantime, insurance channel broadening is adding to the breakdown of identifiable information. Insurance-related exercises should be possible using cell phones separated from the conventional face-to-face and online Insurance [? ? ]. This can be seen as an expansion to data storehouses in the Insurance business. Given more prominent channel enhancement and the development of ranges where fraud can happen, it is vital for insurers to have reachable enterprise-level data about their business and clients.

### 2.3 Data Integration:

Analytics assumes a vital part in incorporating information. Viable fraud recognition abilities can be worked by joining information from different sources. Analytics additionally help in integrating inside information with outsider information that may have predictive significance, for example, public records. Information sources with derogatory properties are on the whole public documents that can be incorporated into a model. Cases include liquidations, liens, criminal records, judgment, abandonment, or even deliver change speed to show transient conduct. Different sorts of outsider information can be useful in upgrading effectiveness, for example, audit evaluating data to decide whether harms coordinate portrayal or misfortune or injury being guaranteed [? ]. A standout amongst the most under-used information sources is doctor's visit expense audit information. This information, if utilized as a part of a model legitimately, is a gold dig for organizations researching medical fraud. Revealing peculiarities, in charging and adding these to the next scoring motors or interpersonal organization analytics will diminish the measure of time an agent or expert spends endeavoring to pull the majority of the pieces together to recognize deceitful action.

## **2.4 Harnessing Unstructured Data:**

Analytics is useful for getting the best incentive from unstructured information. Fraud can be delicate or hard. This depends on whether it comprises of a policyholder's misrepresented cases, or on the off chance that it contains of a policyholder arranging or creating a misfortune. At an abnormal state, fraud can happen amid commission discounting, because of false documentation, an arrangement between parties or from is offering [? ]. Albeit bunches of organized data is put away in an information distribution center as a component of numerous applications, a significant portion of the vital data about a fraud is in unstructured information, for example, outsider reports, which are not assessed. In most insurance firms, data accessible in online networking is not suitably stored. An uncommon investigative-unit specialist will concur that unstructured information is vital for fraud examination. Since textual information is not straightforwardly utilized for reporting, it does not discover a place in most information stockrooms [? ]. This is the place content examination can assume a crucial part in checking on this unstructured information and giving some valuable experiences in fraud discovery.

## **3 RELEVANCE OF BIG DATA IN INSURANCE FRAUD DETECTION**

Big data analytics is a reality for the insurance company because of its capability to enhance various conventional technologies and be used to detect fraudulent acts. In the proceeding section, the relevance of big data and insurance fraud detection will be examined.

### **3.1 Text analysis**

In numerous Insurance fraud recognition ventures, from 33% to oneportion of factors utilized as a part of the fraud location model originate from unstructured content data. This is particularly helpful for long-tail claims, for example, damage claims, because the best information frequently is found in claim notes [? ]. Content mining is something beyond keyword sorting. Excellent content analytics apparatuses translate the importance of the words to establish context. Innovation that is adroit at preparing common dialect can help remove factors from the unstructured content that can be utilized for assist fraud modeling.

### **3.2 Data Management**

Regardless of where your information is stored fi! from legacy frameworks to the valid information stockpiling structure, Hadoop fi! an information administration framework can enable insurers to make reusable information rules. They give a standard, repeatable strategy for enhancing and incorporating information [? ]. Preferably, you need a framework that interfaces with different information sources. It ought to have streamlined information league, relocation, synchronization, organization, and visual assessment.

### **3.3 Event Stream Processing**

This enables insurers to investigate and processes in movement (i.e., process streams). Rather than putting away information and running questions against data, you store the inquiries and stream the data through them [? ]. This is foundational to both ongoing

fraud identification (invigorating fraud scoring) and successful utilization of great high-speed information sources similar to vehicle telematics.

### **3.4 Hadoop**

A free programming structure that assesses and prepares of tremendous collected information in a distributed environment of computing. It offers gigantic details stockpiling and super-quick processing at around 5 percent of the cost of convection less-adaptable databases. Hadoop's mark quality is the capacity to deal with organized and unstructured information (counting sound, text, and visual), and in expansive volumes. Insurers either can employ Hadoop specialists to exploit the structure or purchase items that scaffold to existing databases and information distribution centers[? ? ]. This foundational innovation for making predictive analytics models stays one-step in front of fraudsters and spillage of paid-out cases cash. The exchange observing advancement innovation used to battle regularly complicated illegal tax avoidance utilizes Hadoop as a center stockpiling and sorting out innovation. Complex organized crack rings and therapeutic factories, for instance, are conveying progressively modern techniques for laundering cash stolen from auto insurers.

### **3.5 In memory**

In-memory analytics is a processing style in which all information utilized by an application is put away inside the principal memory of the computing condition. Instead of being available on a disc, the data stays suspended in the mind of useful sets of PCs. Different clients can share this information with numerous applications in a quick, secure, and simultaneous way. In-memory analytics likewise exploits multi-threading and distributed registry [? ? ]. This implies clients can disseminate the information (and complex workloads that process the data) over different machines in a group or inside a single server condition. In-memory analytics manages questions and information analytics, yet also is utilized with more-complex procedures, for example, predictive analytics, machine learning, and analytics. The sorts of neural-network analytics that assist insurer in discovering association among suspects sustaining claim and premium fraud depending on the kind of processes

### **3.6 Software as a Service (SaaS)**

Predictive modeling and different analytics were accessible to large insurance net providers willing to introduce the innovation on location as of not long ago. Software as a service has advanced to even where genuinely little insurers can exploit Big Data analytics [? ]. Insurance providers subscribe to a service keeps running by a seller as opposed to paying for the vast buy, establishment, and support of in-house frameworks. SaaS likewise is named "on-demand software."

## **4 CONCLUSION**

Big data analytics is efficient means that insurance organization can use to structure their data in a manner to detect insurance fraud analyze events. More importantly, big data analytics offers implies that insurance companies can use to develop predictive analytics to identify unknown and suspicious events taking place within

databases systems. More importantly, big data analytics provides means for management of large insurance data efficiently. Additionally, big data analysis can be integrated with another source of information such as public records to determine individual profiles and chances of committing an offense. Notably, big data analytics as SaaS can be used with a different level of insurance firms to detect fraudulent activities in a cost-effective manner.

## **ACKNOWLEDGMENTS**

The authors would like to thank Dr. Gregor von Laszewski for his support and formatting in writing this paper.

## **REFERENCES**

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I couldn't open database file paper2.bib
---line 49 of file report.aux
: \bibdata{paper2}
:
I'm skipping whatever remains of this command
I found no database files---while reading file report.aux
Warning--I didn't find a database entry for "1"
Warning--I didn't find a database entry for "2"
Warning--I didn't find a database entry for "4"
Warning--I didn't find a database entry for "3"
Warning--I didn't find a database entry for "5"
(There were 2 error messages)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

---

```
[2017-11-06 17.35.16] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.1 L33   : [1] undefined
p.1 L33   : [2] undefined
p.1 L36   : [4] undefined
p.1 L39   : [3] undefined
p.1 L45   : [4] undefined
p.1 L45   : [5] undefined
p.1 L49   : [2] undefined
p.1 L53   : [1] undefined
p.1 L53   : [3] undefined
p.2 L61   : [4] undefined
p.2 L67   : [3] undefined
p.2 L72   : [1] undefined
p.2 L75   : [2] undefined
p.2 L75   : [3] undefined
p.2 L78   : [1] undefined
p.2 L78   : [2] undefined
p.2 L81   : [2] undefined
Missing character: ""
Missing character: ""
Empty 'thebibliography' environment.
```

```
There were undefined citations.  
Typesetting of "report.tex" completed in 0.8s.
```

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

```
name: Qiaoyi Liu  
hid: 106  
paper1: Oct 27 17 100%  
paper2: 100%  
project: 0%
```

```
yamlcheck
```

```
wordcount
```

```
3  
wc 106 paper2 3 1721 report.tex  
wc 106 paper2 3 1663 report.pdf  
wc 106 paper2 3 196 report.bib
```

```
find "
```

```
82: Predictive modeling and different analytics were accessible to  
large insurance net providers willing to introduce the innovation  
on location as of not long ago. Software as a service has advanced  
to even where genuinely little insurers can exploit Big Data  
analytics \cite{2}. Insurance providers subscribe to a service  
keeps running by a seller as opposed to paying for the vast buy,  
establishment, and support of in-house frameworks. SaaS likewise  
is named "on-demand software."
```

```
passed: False
```

```
find footnote
```

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

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

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

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

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

```
find textwidth
```

---

```
passed: True
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
I couldn't open database file paper2.bib
---line 49 of file report.aux
: \bibdata{paper2}
:
I'm skipping whatever remains of this command
I found no database files---while reading file report.aux
Warning--I didn't find a database entry for "1"
Warning--I didn't find a database entry for "2"
Warning--I didn't find a database entry for "4"
Warning--I didn't find a database entry for "3"
Warning--I didn't find a database entry for "5"
(There were 2 error messages)
```

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 8212  
non ascii found 8212

---

The following tests are optional

---

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Application in Amazon

Shiqi Shen

Indiana University Bloomington

1575 S Ira St

Bloomington, Indiana 47401

shiqshen@indiana.edu

## ABSTRACT

This case study evaluates being data application in Amazon, with specific focus on how the company has employed big data to enhance its performance. Amazon has created robust applications from big data that has enabled it to give customers more targeted product recommendations and enhance the quality of care for them. The all-rounded customer profiles created using big data resources has enabled the firm to send customized and personalized marketing messages for its customers. Other firms have also adopted the big data tools offered by Amazon to enhance their performance and revenue flows. The other parts of the paper evaluate how amazon used big data to enhance operations and improve its performance.

## 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 firm's 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 customer's needs and tastes. Since it was founded, Amazon has created a name for itself because of its superior product recommender system, which suggests products to consumers on the basis of their last purchase. The major driving force behind the recommender system is the data gathered from the customers. The product recommender system is essential for the personalization of each customer's experience when they are shopping in the firm's online store [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 firm's shopping portal, availability of the product, competitor's 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 firm in 2009, the procedures were integrated together with those of Amazon. Today's business environment is changing at a rapid rate, and consumers are also using their voices faster. Within a few moments after undergoing a bad experience, customers can swiftly move into social media and spread the news about their negative experience [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 firm's revenues have increased year after year. The significance of this application pushed the company to patent it to prevent other companies from using it without authorization. Reorganizing the purchase process is currently one of the most significant differentiates in the current marketplace. The service enables users to make payments without having to exchange cards or money physically. Amazon has also greatly benefited from impulse buying, which is accelerated by one-click buying. Research has shown that the largest percentage of people normally purchase things they don't require or did not plan to purchase in the first place [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 firm's products in patterns similar to Amazon. That is to say that the retailers will also be able to stalk their customers, recommend products to them, and improve their customer experience.

### **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 serviceless, 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.amazon.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](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.

## bibtex report

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

## bibtext \_ label error

bibtext space label error

bibtext comma label error

# latex report

[2017-11-06 17.35.21] 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 0.9s.
```

```
=====
```

## Compliance Report

```
=====
```

```
name: Shiqi Shen
hid: 109
paper1: complete 100% Oct 27th
paper2: complete 100% Oct 4th
project: 0%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
3
wc 109 paper2 3 2060 report.tex
wc 109 paper2 3 2105 report.pdf
wc 109 paper2 3 199 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

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

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

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

```
find ""
```

---

```
passed: True
```

```
ascii
```

---

```
non ascii found 8217
```

---

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

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

```
find newline
```

---

```
passed: True
```

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

```
find cite {
```

---

```
passed: True
```

# Using MQTT for Communication in IoT Applications

Arnav

Indiana University Bloomington  
Bloomington, Indiana 47408  
aarnav@iu.edu

## ABSTRACT

With the increase in the number of edge devices and their applications in settings like sensor networks, it is crucial to allow communication between the sensing devices and actuators, which may not be directly connected. To allow services built on various different platforms and using different hardware to communicate, a data agnostic, fast and reliable mechanism is needed to allow communication between these devices. In addition to communication, the data generated by these devices must be analyzed and security of this data is highly important. MQTT is a common, easy to use, queuing protocol that helps meet these requirements.

## KEYWORDS

i523, HID201, MQTT, IoT, Edge Computing, Security

## 1 INTRODUCTION

As Internet of Things (IoT) applications and sensor networks become commonplace and more and more devices are being connected, there is a need to allow these devices to communicate. It is often the case that these edge devices have a very limited memory and need to conserve power, and may not have enough computing capacity to process traditional HTTP web requests efficiently [2][10]. Monitoring the state of a remotely located sensor using HTTP would require sending requests and receiving responses to and from the device frequently, which may not be efficient on small chips on these sensors [2].

Message Queue Telemetry Transport (MQTT) is a lightweight machine to machine (M2M) messaging protocol, based on a client/server based publish-subscribe model, that is an elegant solution for such scenarios. MQTT was first developed in 1999 by Andy Stanford-Clark and Arlen Nipper to connect oil pipelines [10]. The protocol has been designed to be used on top of TCP/IP protocol in situations where network bandwidth, and available memory are limited allowing low power usage. It allows efficient transmission of data to various devices listening for the same event, and is scalable as the number of devices increase [23][15].

“Eclipse Paho project, maintained by eclipse iot, MQTT clients for various languages such as C, Python and Lua. It also includes an open source Mosquitto broker and a Really Small Message Broker from IBM” [15][5].

## 2 MQTT DETAILS

MQTT works on a publish-subscribe model, which contains 3 entities, a publisher, that sends a message, a broker, that maintains queue of all messages based on topics and multiple subscribers that subscribe to various topics they are interested in [17].

Gregor von Laszewski

Indiana University  
Smith Research Center  
Bloomington, IN 47408, USA  
laszewski@gmail.com

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 other's identity but only that of the broker and the publisher and the subscribers do not have to be running at the same time [12].

### 2.1 Topics

MQTT uses a cleverly designed hierarchy of topics, which are related to all messages. These topics are recognised by strings separated by a forward-slash (/) and each part representing a different topic level.

For example *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 allows subscribers to filter what messages to receive based on the topic hierarchy. Thus, when a publisher publishes a message related to a topic to the broker, the message is forwarded to all the clients that have subscribed to the topic of the message or a topic that has a lower depth of hierarchy [12] [17].

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 [12]. 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 [23].

### 2.2 Callbacks

One of the main advantages of using MQTT is that it allows asynchronous behaviour with the help of callbacks. Both the publisher and subscriber do not have to wait to publish a message or receive one, and can perform other tasks in a non-blocking manner [12] [6].

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].

### 2.3 Quality of Service

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) [13][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 [13] [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 3 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 [13] [6].

The various levels of quality of service allow the use of this protocol in a variety of applications.

### 3 SECURITY WITH MQTT

The MQTT specification uses TCP/IP to deliver the message to the subscribers, but it does not provide any form of security by default to make it useful for resource constrained IoT devices. “It allows the use of username and password for authentication, but by default this information is sent as plain text over the network, making it susceptible to man-in-the-middle attacks” [16] [14]. 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 [14].

#### 3.1 Using TLS/SSL

Transport Layer Security (TLS) and Secure Sockets Layer (SSL) are cryptographic protocols that establish the identity of the server and client with the help of a handshake mechanism which uses trust certificates to establish identities before encrypted communication can take place [4]. If the handshake is not completed for some reason, the connection is not established and no messages are exchanged [14]. “Most MQTT brokers provide an option to use TLS instead of plain TCP and port 8883 has been standardized for secured MQTT connections” [16].

Using TLS/SSL security however comes at an additional cost. If the connections are shortlived 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 shortlived, 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 [16][14].

#### 3.2 Using OAuth

OAuth is an open protocol that allows access to a resource without providing unencrypted credentials to the third party. Although MQTT protocol itself does not include authorization, many MQTT brokers include authorization as an additional feature [4]. OAuth2.0 uses JSON Web Tokens which contain information about the token and the user and are signed by a trusted authorisation 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 authorisations for the client [9]. Since there can be a large number of clients and it can become impractical to authorise everyone, clients may be grouped and the authorizations may be checked for each group [4].

## 4 INTEGRATION WITH OTHER SERVICES

As the individual IoT devices perform their respective functions in the sensor network, a lot of data is generated which needs to be processed. 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 [21] to get real time data streams and allow analytics and online machine learning in a fault tolerant manner [24].

ELK stack (elastic-search, logstash and kibana) is an open-source project designed for scalability which contains three main software packages, the *elastic-search* search and analytics engine, *logstash* which is a data collection pipeline and *kibana* which is a visualization dashboard [7]. Data from an IoT network can be collected, analysed and visualized easily with the help of the ELK stack as shown in [20] and [19].

MQTT broker services can be utilised 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].

## 5 USE CASE

MQTT can be used in a variety of applications. This section explores a particular use case of the protocol. A small network was set up with three devices to simulate an IoT environment, and actuators were controlled with the help of messages communicated over MQTT.

### 5.1 Requirements and Setup

The setup used three different machines. A laptop or a desktop running the MQTT broker, and two raspberry pis configured with raspbian operating system. Eclipse Paho MQTT client was setup on each of the raspberry pis [6]. Additionally all three devices were connected to an isolated local network.

Grovepi shields for the raspberry pis, designed by Dexter Industries were used on each of the raspberry pis to connect the actuators as they allow easy connections of the raspberry pi board [11]. The actuators used were Grove relays [22] and Grove LEDs [18] which respond to the messages received via MQTT.

To control the LEDs and relays, the python library cloudmesh.pi [3], developed at Indiana University was used. The library consists of interfaces for various IoT sensors and actuators and can be easily used with the grove modules.

## 5.2 Results

The two raspberry pis subscribe connect to the broker and subscribe with different topics. The raspberry pis wait for any messages from the broker. A publisher program that connects to the broker publishes messages to the broker for the topics that the two raspberry pis had registered. Each raspberry pi receives the corresponding message and turns the LEDs or relays on or off as per the message.

On a local network this process happens in near real time and no delays were observed. Eclipse iot MQTT broker ([iot.eclipse.org](http://iot.eclipse.org)) was also tried which also did not result in any significant delays.

Thus it is observed that two raspberry pis can be easily controlled using MQTT. This system can be extended to include arbitrary number of raspberry pis and other devices that subscribe to the broker. If a device fails, or the connection from one device is broken, other devices are not affected and continue to perform the same.

This project can be extended to include various other kinds of sensors and actuators. The actuators may subscribe to topics to which various sensors publish their data and respond accordingly. The data of these sensors can be captured with the help of a data collector which may itself be a different subscriber, that performs analytics or visualizations on this data.

## 6 CONCLUSION

We see that as the number of connected devices increases and their applications become commonplace, MQTT allows different devices to communicate with each other in a data agnostic manner. MQTT uses a publish-subscribe model and allows various levels of quality of service requirements to be fulfilled. Although MQTT does not provide data security by default, most brokers allow the use of TLS/SSL to encrypt the data. Additional features may be provided by the broker to include authorization services. MQTT can be easily integrated with other services to allow collection and analysis of data. A small environment was simulated that used MQTT broker and clients running on raspberry pis to control actuators

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper, and for providing all the hardware needed to complete the setup at smith research center.

The authors would also like to thank the Associate Instructors of the class for answering questions regarding the paper on piazza, which helped all students in the class.

## REFERENCES

- [1] apache. [n. d.]. apache storm. apache storm website. ([n. d.]). <http://storm.apache.org/>
- [2] Paul Caponetti. 2017. Why MQTT is the Protocol of Choice for the IoT. xively.com blog website. (august 2017). <http://blog.xively.com/why-mqtt-is-the-protocol-of-choice-for-the-iot/>
- [3] cloudmesh. 2017. cloudmesh.pi. github. (october 2017). <https://github.com/cloudmesh/cloudmesh.pi>
- [4] Ian Craggs. 2013. MQTT security: Who are you? Can you prove it? What can you do? IBM developer works website. (march 2013). [https://www.ibm.com/developerworks/community/blogs/c565c720-fe84-4f63-873f-607d87787327/entry/mqtt\\_security?lang=en](https://www.ibm.com/developerworks/community/blogs/c565c720-fe84-4f63-873f-607d87787327/entry/mqtt_security?lang=en)
- [5] eclipse. [n. d.]. mqtt broker. eclipse mosquitto website. ([n. d.]). <https://mosquitto.org/>
- [6] eclipse paho. [n. d.]. Python Client - documentation. eclipse paho website. ([n. d.]). <https://www.eclipse.org/paho/clients/python/docs/>
- [7] elastic.io. [n. d.]. ELK stack. elastic.io website. ([n. d.]). <https://www.elastic.co/products>
- [8] erlang mqtt. [n. d.]. erlang mqtt broker. wmqtt website. ([n. d.]). <http://emqttd.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/hive-mq-security-fundamentals-oauth-2-0-mqtt>
- [10] hivemq. [n. d.]. intrewebsite mqtt. hivemq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-essentials-part-1-introducing-mqtt>
- [11] Dexter Industries. 2017. GrovePi. Dexter Industries website. (2017). <http://www.dexterindustries.com/grovepi/>
- [12] Hive mq. [n. d.]. MQTT Essentials Part 2: Publish & Subscribe. HiveMQ website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-essentials-part2-publish-subscribe>
- [13] 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>
- [14] Hive Mq. [n. d.]. MQTT Security Fundamentals: TLS / SSL. hive mq website. ([n. d.]). <https://www.hivemq.com/blog/mqtt-security-fundamentals-tls-ssl>
- [15] Mqtt. [n. d.]. Mqtt official website. mqtt official website. ([n. d.]). <http://mqtt.org/>
- [16] 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/>
- [17] 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/>
- [18] seed studio. 2017. Grove LED socket kit. Seed studio website. (October 2017). [http://wiki.seedcc/Grove-LED\\_Socket\\_Kit/](http://wiki.seedcc/Grove-LED_Socket_Kit/)
- [19] 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/>
- [20] 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/>
- [21] Apache storm. [n. d.]. Storm MQTT Integration. Apache storm website. ([n. d.]). <http://storm.apache.org/releases/1.1.0/storm-mqtt.html>
- [22] Seed Studio. 2017. Grove Relay. seed studio website. (October 2017). <http://wiki.seedcc/Grove-Relay/>
- [23] 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].
- [24] Wikipedia. 2017. Storm (event processor) – Wikipedia, The Free Encyclopedia. (2017). [https://en.wikipedia.org/w/index.php?title=Storm\\_\(event\\_processor\)&oldid=80871136](https://en.wikipedia.org/w/index.php?title=Storm_(event_processor)&oldid=80871136) [Online; accessed 6-November-2017].

## 7 ISSUES

DONE:

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

### 7.1 Assignment Submission Issues

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

### 7.2 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

### 7.3 Formatting

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

Other formatting issues

#### 7.4 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}

#### 7.5 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

#### 7.6 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

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 inflate your paper if you are below the page limit

#### 7.8 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 float. 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

#### 7.7 Structural Issues

Acknowledgement section missing

Incorrect README file

```
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 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 13 warnings)
```

```
bibtext _ label error
```

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

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

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

```
[2017-11-06 17.35.26] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
```

```
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 1.2s.
./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)
```

---

## Compliance Report

---

```
name: Arnav, Arnav
hid: 201
paper1: 20th Oct 2017 100%
paper2: 80%
project: not started
```

```
yamlcheck
```

---

```
wordcount
```

---

```
(null)
wc 201 paper2 (null) 2225 report.tex
wc 201 paper2 (null) 3135 report.pdf
wc 201 paper2 (null) 873 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
figures 0
```

```
tables 0
```

```
includegraphics 0
```

```
labels 0
```

```
refs 0
```

```
floats 0
```

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

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

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

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

```
Label/ref check
```

```
passed: True
```

```
When using figures use columnwidth
```

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

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

```
find textwidth
```

---

```
passed: True
```

below\_check

---

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty year in 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 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 13 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
```

# Algorithms for Big Data Analysis

Jyothi Pranavi Devineni  
Indiana University Bloomington  
Bloomington, Indiana  
jyodevin@umail.iu.edu

## ABSTRACT

Analysis of data is not easy, especially when the data is unstructured or has many features. However, there are many algorithms for data pre-processing, feature selection, classification, regression, etc. These algorithms make it simpler to understand and analyze the data. One of the main types of algorithms for analyzing the data is clustering algorithms. They help to categorize the data into clusters. All the related data is collected under one cluster. Clustering facilitates easy analysis of data. However, when it comes to big data, ordinary clustering algorithms might not work because of the absence of formal categorization. Hence, there are modified clustering algorithms for big data and they are comparable to the already existing algorithms for ordinary data.

## KEYWORDS

Clustering, Big Data, Fuzzy, Partitioning

## 1 INTRODUCTION

With the advances in technology, social media, search engines and other online websites like online shopping, became a part and parcel of everyone's life. With this, there are massive amounts of data available today which can be used for many applications such as improving the sales, predicting the future outcomes, etc. However, the amount of data produced by such websites and multinational companies is enormous. Conventional databases cannot store data that huge. Also, processing of such massive amounts of data is challenging. There are frameworks like Hadoop and its ecosystems make it easier to manage big data. Another efficient method of dealing with big data is to cluster the data without making it losing the information. There are effective clustering algorithms for big data which aim at producing such informative clusters which can be used by common people as well as corporate world.

When it comes to Big Data, it is important to address three Vs. The first and the most important is the "Volume" of the data. To deal with huge volumes of data, a change in the storage architectures is required. Hadoop databases like HDFS and HBase can be used to store large volumes of data. Having dealt with the volume of big data, the next important feature is the "Velocity" of data. Data is generated as a continuous flow from online websites and social media sites. Hence, such data should be processed dynamically without much time lapse. The last feature to be addressed is the "Variety" of the data.

Different types of data such as images, text, audio, etc are produced by companies and online websites. This data may be structured, semi-structured or unstructured. The proposed clustering algorithms for big data must be able to take care of these three features. In other words, according to our requirement, a suitable

clustering algorithm should be used. Although there are many clustering algorithms for machine learning[6], data mining[3], wireless signal processing[1] and so on, it is not obvious which algorithm to use for a given data. It is the work of the researcher to carefully choose among the available algorithms.

## 2 CLUSTERING CRITERION

In order to consider a clustering algorithm for clustering big data, the algorithm has to address the three Vs of big data. A clustering algorithm for huge volumes of data should consider the size of the data and must be able to handle the high dimensionality of the data and outliers.

Similarly, when working with wide variety of data, to select an appropriate clustering algorithm, the factors to be considered are the type of the data set and size of the cluster. To select a clustering algorithm for data being generated continuously or with high velocity, the runtime of the algorithm is of utmost importance and so is the complexity of the algorithm.

The features to be considered while looking for an appropriate clustering algorithm can be summarized as follows[5]:

- (1) **Size of the data:** The size of the data is a major concern when it comes to applying normal clustering algorithms to big data. Clustering algorithms which work very efficiently for small data sets might not work well for big data.
- (2) **Handling High Dimensionality:** When trying to cluster huge volumes of data, it is important to take into account many or all of the attributes or features of data into consideration in order to get maximum possible information from the data. There are methods for dimensionality reduction to keep the most important features of the data and discard the rest whose presence or absence doesn't affect the analysis much. As the dimensionality increases, the data becomes sparse and clustering becomes difficult.
- (3) **Handling the Outliers:** When clustering the data, there might be some data points which are left out as we cannot include them in any cluster. Such data points, which do not conform to the properties of any of the designed clusters by the algorithm are called outliers or noisy data in other words. Hence, a clustering algorithm must be capable of handling the noisy data, by not losing the informative data.
- (4) **Type of Data:** Conventional clustering algorithms are designed for either numeric data or categorical data. But, in the real world, the data is available as numeric, categorical and also a mix of both. Hence clustering algorithms designed for numeric and categorical data does not work on mixed data.
- (5) **Shape of the Cluster:** An efficient clustering algorithm should be able to handle different data, which produces clusters of different shapes.

- (6) **Time Complexity or Run time of the Algorithm:** The clustering algorithms perform merely efficiently when they have used for clustering again and again to obtain the final clusters with good accuracy. Hence, if the runtime of the algorithm is too long, it takes infinitely long time to obtain the required clusters, especially while dealing with big data. Hence, the algorithm should be able to run within a finite time.
- (7) **Veracity:** An efficient clustering algorithm must be capable of producing the same data clusters, irrespective of the order in which the data is given.

### 3 TYPES OF CLUSTERING ALGORITHMS

Their clustering algorithms can be categorized based on the method of clustering they follow as follows:

- (1) Partitioning-Based
- (2) Hierarchical-based
- (3) Density-based
- (4) Model-based

#### 3.1 Partitioning-Based

In partitioning-based clustering algorithms, the data is divided into distinctive partitions. Each partition represents a cluster. The clusters should satisfy two characteristics: (i) Each cluster should contain at least one data point or object and (ii) Each data point should belong to only one of the clusters at any given a point of time. Initially, the data points are partitioned based on a union. For example, in K-Means algorithm, the center is the arithmetic mean of all data points belonging to a cluster and the cluster is represented by the arithmetic mean. K-Medoids, K-modes, FCM, and CLARA are other examples of partition based clustering algorithms.

*3.1.1 Fuzzy-C-means(FCM).* Fuzzy-C-means is a fuzzy clustering algorithm which is based on K-means[2]. It is a soft clustering algorithm which places each data point in one or more clusters, with some degree of belief. The degree of belief of a data point ranges between 0 and 1 and according to the fuzzy rule, the sum of the degree of beliefs for a given data point over all clusters should be equal to 1. The fuzzy clustering finds the center of the cluster and updates the data points and their degrees of membership. This algorithm has the same drawback as that of the K-Means algorithm, i.e, the final clusters obtained are based on the selection of the initial weights as that of K-means and also the centers are local to that specific cluster.

#### 3.2 Hierarchical-based

In this type of clustering, data is organized in a hierarchical fashion and a single cluster may be divided into a number of clusters as the hierarchy progresses. The clustering can be agglomerative or divisive. As the name suggests, in agglomerative clustering, each object is treated as a cluster and as the time progresses, two or more related objects merge into one cluster. Alternatively, in divisive clustering, the whole data set is considered as one cluster and as the time progresses, the cluster is divided into different clusters based on common properties. This process continues in both agglomerative and divisive clustering techniques until a desired number

of clusters are reached. Algorithms like BIRCH, Chameleon, and CURE are the examples of hierarchical-bases clustering.

*3.2.1 BIRCH.* The BIRCH algorithm[7] builds a CF tree or clustering feature tree by scanning the data dynamically. It initially scans the data and constructs an in-memory CF tree and then runs the algorithm to determine the clustering leaf nodes. It also assumes a branching factor B and a threshold T initially. The CF tree is constructed with the assumed branching factor and the clusters are created with a diameter within the threshold. The clusters created are hence circular. Whenever a data point is encountered, the algorithm traverses from the root node of the tree to the nearest child until a leaf node cluster is reached. Once the leaf node is reached, the data point is tested if it belongs to that cluster and if not, a new cluster with a diameter greater than the current threshold is created. This algorithm can deal with the noisy data effectively but cannot deal with clusters of different shapes, as the clusters created in this algorithm are spherical in shape. Also, for the different order of the data points given, different clusters are generated. BIRCH works effectively with one scan of the data but can become more efficient if the data is scanned repeatedly.

#### 3.3 Density-based

In density-based clustering, data points are clustered based on the density. The cluster progresses in the direction of the density, hence density clustering produces clusters of different shapes. This type of clustering is capable of handling the outliers or noisy data. Algorithms like DENCLUE and OPTICS are examples of density-based clustering algorithms.

#### 3.4 Model-based

Model-based clustering algorithms assume that the data is generated by some probabilistic distribution and generate a fixed number of robust clusters, determined by some statistics such as the log likelihood. There are two types of model-based approaches, statistical and neural networks. In statistical approach determines the clusters based on the probabilities, whereas in neural network approach, the data points are represented as a series of connected input/output units, called perceptrons and the connections between them are assigned specific weights. The neural networks are famous for clustering as they can perform parallel processing and also they can adjust their weights according to the errors propagated using backpropagation. Expectation Maximization and COBWEB are the examples of model-based clustering algorithms.

*3.4.1 Expectation-Maximization (EM).* There are three major steps in the EM algorithm. They are Initialization, Expectation, and Maximization. Its goal is to find a maximum likelihood solution.[4] It assumes that the data points are statistically distributed. In the initialization step, it assumes a certain number of clusters and also the respective means and variances for each distribution and the prior probabilities of the clusters, which should sum to one. In the expectation step, the posterior probability of each data point belonging to a particular cluster is calculated. In the maximization step, the algorithm tries to maximize the expectation. The new means, variances and prior probabilities are calculated. The expectation and maximization steps are performed iteratively, maximizing the

likelihood. The algorithm always converges but is prone to arrive at local maxima.

## 4 CONCLUSIONS

Clustering algorithms have been used whenever it comes to data analysis. But when it comes to big data, conventional clustering algorithms do not work and there is a need to follow specific algorithms which can handle the volume, variety, and velocity of big data. Each of the clustering algorithms can be used under different requirements as stated.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor Von Laszewski and all the associate instructors of the course I-523 for guiding us through.

## REFERENCES

- [1] A. A. Abbasi and M. Younis. 2007. "A survey on clustering algorithms for wireless sensor networks". *IEEE* 30, 14 (Oct. 2007), 2826–2841.
- [2] J. C. Bezdek, R. Ehrlich, and W. Full. 1984. "FCM: The fuzzy c -means clustering algorithm" *Computers & Geosciences* 10, 2 (1984), 191–203.
- [3] C. Zhai C. C. Aggarwal. 2012. "A survey of text clustering algorithms". *Mining Text Data* 2, 3 (Jan. 2012), 77–128.
- [4] "A. P. Dempster, N. M. Laird, and D. B. Rubin". 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B* 39, 1 (1977), 1–38.
- [5] Adil Fahad, Najlaa Alshatri, and Zahir Tari. 2014. "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis". *IEEE* 2, 3 (Sept. 2014), 267 – 279.
- [6] R. Xu and D. Wunsch. 2005. "Survey of clustering algorithms". *IEEE* 16, 3 (May 2005), 645–678.
- [7] T. Zhang, R. Ramakrishnan, and M. Livny. 1996. "BIRCH: An efficient data clustering method for very large databases". *ACM* 25, 2 (June 1996), 103–114.

# Machine Learning Optimization for Big Data

Ajinkya Khamkar  
Indiana University  
Bloomington, IN 47408, USA  
adkhamka@iu.edu

## ABSTRACT

The last decade has seen the rise of big data. Industries and organizations collect consumer and machinery data to make data driven business decisions. Traditional naive variants of machine learning algorithms are ill equipped to handle the challenges posed by big data. Significant alterations are required to existing algorithms to ensure optimality and efficiency in big data applications.

## KEYWORDS

Machine Learning, Optimization, I523, HID211

## 1 INTRODUCTION

The last decade has seen the rise of big data. Industries and organizations collect consumer and machinery data to make data driven business decisions. Machine learning techniques are used to drive data driven decisions in organizations. Traditional machine learning algorithms were designed prior to the advent of big data era. They are ill-equipped for handling the scale and volume of big data tasks [6]. Recent advancements in hardware allow for running machine learning algorithms in parallel. In a parallel environment these algorithms suffer asynchronous gradient updates. In section 2, we discuss the need for efficient algorithms. In section 3 we discuss traditional machine learning algorithms and their drawbacks. In section 4, we discuss improvements to existing methods to support big data tasks. In section 5, we discuss various techniques to deploy these algorithms in a parallel environment for efficient and optimal performance. In section 6, we conclude our discussion.

## 2 DATA

Multinational Corporations and Organizations collect consumer information in order of terabytes. Social Media Platforms are regularly queried by millions of users from all across the globe. E-commerce websites process hundred thousands of orders daily. Sensors for varied tasks collect information per fraction of a second. This arises the need for computationally efficient algorithms to process and convert this data into information.

## 3 TRADITIONAL ALGORITHMS

Machine learning algorithms can broadly be classified in to supervised [3] and unsupervised approaches. Supervised approaches require a training phase to train the parameters of the algorithm to draw decision boundaries. Unsupervised approaches require reconfiguration of the decision boundaries for a new batch of input data. Further, these algorithms can be characterized by their ability to draw linear and non linear decision boundaries. Non linear decision boundaries are difficult to draw as they require computation of higher order polynomials to best fit the input data. These

decision boundaries are estimated using parameters of the algorithm. Traditional machine learning algorithms rely on gradient descent to estimate the true parameters representing the underlying data distribution[1]. Gradient descent seeks to iteratively optimize parameters such that they minimize the given error function.

### 3.1 Drawbacks of Traditional Algorithms

Big data is highly unconstrained and can span over billions of records and thousands of parameters to choose from. Data is pulled from a variety of sources and collected into data warehouses. With increasing dimensionality the model runs into following problems.

- The complexity of model increases. The decision boundary can span across multiple dimensions making it difficult to comprehend the impact of features.
- The variance of model increases. This leads to over fitting. The model will tightly fit to the input data and fail to generalize for unseen test data.
- Leads to wastage of computing resources. Resources would be spent on computing errors or coefficients of features which contribute little to the decision boundary.

Directly training machine learning algorithms to draw decision boundaries on this data is highly inefficient [1]. Traditional machine learning algorithms use gradient descent algorithm to compute the parameters. Classification and regression tasks can be formulated as an optimization task and parameters can be tuned to minimize the generated error. Error is computed during the forward pass of the algorithm. Gradient of the parameters  $x$  is the partial differentiation of the error with respect to parameters.

$$\nabla = \left( \frac{\partial f}{\partial x_1} \quad \dots \quad \frac{\partial f}{\partial x_n} \right)$$

Many algorithms compute the *Hessian* of the error for smoother transitions over the error surface. *Hessian* is the second order derivative of the error function.

$$\nabla^2 = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

At higher dimensions it becomes infeasible to compute the true error gradient. We are thus required to compute an approximated error gradient. There is a trade off between convergence to the true gradient and computation time. In the following section we will discuss multiple gradient approximation techniques. We also discuss their ability to scale and outperform traditional gradient descent techniques for big data tasks. Training these algorithms on commodity hardware pose additional space and computation constraints. The sheer volume of the data ensures it cannot be stored and retrieved from single machines. Data is required to

be distributed across several machines and several copies of the algorithm can be trained in parallel to improve efficiency and computation. The major drawback in training algorithms in parallel is asynchronous gradient updates. We discuss various methods to train algorithms in parallel optimally and efficiently.

## 4 IMPROVEMENTS TO TRADITIONAL ALGORITHMS

In this section we will review multiple methods which allow us to approximate the error gradient efficiently. These methods can be scaled to handle big data tasks. These methods converge to the true gradient for sufficiently large number of iterations.

### 4.1 Coordinate Descent Optimization

Coordinate descent algorithms [8] is a derivative-free optimization technique. to approximate the convex function  $f(x)$ , optimize one column of  $f(x)$  using  $\min_{x \in R^n} f(x)$  at every iteration The algorithm converges for strictly convex error surfaces. Algorithm 1 is the general framework of coordinate descent algorithm.

---

#### Algorithm 1 coordinate descent

---

```

1: initialize  $x_0$ 
2: for  $t \in 1...n$  do
3:   Pick coordinate  $i \in 1....n$ 
4:    $x_i^{t+1} = x_i - \lambda[\nabla^t f(x_i)]$ 
```

---

Here  $\lambda$  represents the step size. The algorithm is simple and easily scalable. Block variants of coordinate descent algorithm are discussed below.

Cyclic coordinate descent cycles through each block and computes the descent for each block iteratively. Random Block Coordinate algorithm [5] presented in algorithm 2 draws from a random distribution and updates the parameters. Block Coordinate descent with Gauss-Southwell [5] rule presented in algorithm 3 selects the block which minimizes the error in a greedy manner.

---

#### Algorithm 2 randomized coordinate descent

---

```

1: initialize  $x_0$ 
2: for  $t \in 1...n$  do
3:   sample from block  $i \in 1....n$ 
4:    $x_i^{t+1} = x_i - \lambda[\nabla^t f(x_i)]$ 
```

---



---

#### Algorithm 3 gauss-southwell coordinate descent

---

```

1: initialize  $x_0$ 
2: for  $t \in 1...n$  do
3:   select  $i = \text{argmax}(\nabla^t f(x_i))$ 
4:    $x_i^{t+1} = x_i - \lambda[\nabla^t f(x_i)]$ 
```

---

The major drawback of Coordinate Descent algorithm is it converges for strictly convex optimization. For non-smooth convex optimization we can approximate the non-smoothness using a smooth function prior to performing coordinate descent.

## 4.2 Stochastic Gradient Descent Optimization

Computing the full gradient every iteration is infeasible for big data tasks. Stochastic gradient Descent [1] iteratively computes the gradient per sample in the dataset. Samples are drawn at random from the dataset. This algorithm has 2 major drawbacks.

- As gradients are computed per sample. This leads to high bias and unstable learning, due to uncontrolled gradient jumps
- This method works relatively well for small to medium datasets and remains infeasible for big data

Instead of computing gradient per sample, splitting the dataset into multiple mini batches [1] and sampling randomly or cyclically from the mini-batches as presented in algorithm 4 leads to much stable learning and faster convergence

---

#### Algorithm 4 minibatch stochastic gradient descent

---

```

1: initialize  $w$  and learning rate  $l$ 
2: while no convergence do
3:   Randomly sample from minibatch distribution  $\epsilon$ 
4:   update  $w_{t+1} = w_t - l \frac{1}{|S_k|} \sum \nabla_w f(S_k)$ 
5:   where  $S_k$  is sampled minibatch
```

---

Stochastic gradient descent algorithm is prone to be stuck in local minimum. Convergence can be accelerated using Momentum techniques such as Nestevrov [2], ADAGRAD [2] and ADADELTA [9].

In the following section we discuss implementation of the above algorithms in a parallel computing environment

## 5 PARALLEL IMPLEMENTATION OF GRADIENT DESCENT

Zinkevich, Weimer, Smola & Li, 2010 [10] introduced parallel stochastic gradient optimization technique. This technique is shown to converge and is simple to implement. Gradients generated by the workers in the network are averaged. Algorithm 5 is applied iteratively until convergence

---

#### Algorithm 5 Parallel SGD ( $\{c^1, \dots, c^m\}, T, n, w_0, k$ )

---

```

1: while no convergence do
2:   for machine  $\in \{1\dots,k\}$  in parallel do
3:      $v_i = SGD(\{c^1, \dots, c^k\}, T, n, w_0)$ 
4:      $\nabla v = \frac{1}{k} \sum_{i=1}^k v_i$ 
5:      $v^{+1} = v - \lambda \nabla v$ 
```

---

Meng et al. 2016 [4], introduce an asynchronous stochastic gradient descent variant with stochastic coordinate sampling. They use the following setup. Distributed environment with a master node and  $p$  worker nodes, the parameters  $\theta$  are distributed across several machines and each worker machine has non-overlapping subset of the data.

- (1) Each worker requests for updated parameters from master  $\theta_k$
- (2) Each worker draws a random mini-batch sample  $S_k$  and draws a random set of coordinates  $x_k \subset \theta$

- (3) Each worker computes gradients without synchronization  $\nabla x_k$
- (4) Each worker forwards computed gradient and the sampled coordinates back to the master  $\{\nabla x_k, x_k\}$
- (5) Master updates the parameters asynchronously.

Richtarik and Takac (2012) [7], present a parallel stochastic coordinate descent algorithm where each processor updates a randomly selected subset of coordinates simultaneously.

## 6 CONCLUSION

Machine learning algorithms play an important role for big data applications. We wished to highlight theoretical optimization constraints for big data using traditional techniques. The sheer volume of the data leads to other optimization problems. These include feature reduction, Randomized methods for matrix decomposition over Principal Component analysis and iterative model building which are beyond the scope of this discussion. We presented the drawbacks of traditional gradient descent, which remains the backbone of machine learning algorithms. We discussed several methods to approximate the true gradient of the error function or perform gradient-free parameter updates. We also discussed several parallel implementations of the above techniques for handling big data tasks efficiently and optimally.

## 7 ACKNOWLEDGEMENT

The author would Like to thank Professor Gregor von Laszewski and the Teaching Assistants for their help and guidance.

## REFERENCES

- [1] Léon Bottou. 2010. *Large-Scale Machine Learning with Stochastic Gradient Descent*. Physica-Verlag HD, Heidelberg, 177–186. [https://doi.org/10.1007/978-3-7908-2604-3\\_16](https://doi.org/10.1007/978-3-7908-2604-3_16)
- [2] John Duchi, Elad Hazan, and Yoram Singer. 2010. *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*. Technical Report UCB/EECS-2010-24. EECS Department, University of California, Berkeley. <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html>
- [3] S. B. Kotsiantis. 2007. Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 3–24. <http://dl.acm.org/citation.cfm?id=1566770.1566773>
- [4] Qi Meng, Wei Chen, Jingcheng Yu, Taifeng Wang, Zhi-Ming Ma, and Tie-Yan Liu. 2016. Asynchronous Accelerated Stochastic Gradient Descent. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, USA, 1853–1859. <http://dl.acm.org/citation.cfm?id=3060832.3060880>
- [5] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. 2015. Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Francis Bach and David Blei (Eds.), Vol. 37. PMLR, Lille, France, 1632–1641. <http://proceedings.mlr.press/v37/nutini15.html>
- [6] George Papamakarios. 2014. Comparison of Modern Stochastic Optimization Algorithms. (2014).
- [7] P. Richtárik and M. Takáč. 2012. Parallel Coordinate Descent Methods for Big Data Optimization. *ArXiv e-prints* (Dec. 2012). arXiv:math.OC/1212.0873
- [8] Stephen J. Wright. 2015. Coordinate Descent Algorithms. *Math. Program.* 151, 1 (June 2015), 3–34. <https://doi.org/10.1007/s10107-015-0892-3>
- [9] Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR* abs/1212.5701 (2012). arXiv:1212.5701 <http://arxiv.org/abs/1212.5701>
- [10] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. 2010. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.). Curran Associates, Inc., 2595–2603. <http://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf>

## A ISSUES

DONE:

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

### A.1 Bibliography Errors

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

DONE:

Citations are showing up in the report

### A.2 Formatting

HID not included in the keywords

DONE:

HID Included

### A.3 Writing Errors

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

DONE:

I changed to we

Error in title capitalization

DONE:

Capitalization

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

DONE:

Updated

### A.4 Citation Issues and Plagiarism

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

Lack of citations in Introduction section and also in section 4

DONE:

Done, Citations in introduction and section 4

### A.5 Structural Issues

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

### A.6 Details about the Figures and Tables

Figures, tables and algorithms should be referred to in the text

DONE:

Algorithm refered

re

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--numpages field, but no articleno or eid field, in Kotsiantis
Warning--numpages field, but no articleno or eid field, in Meng
Warning--no number and no volume in 2012arXiv
Warning--page numbers missing in both pages and numpages fields in 2012arXiv
Warning--numpages field, but no articleno or eid field, in Wright
Warning--page numbers missing in both pages and numpages fields in DBLP
Warning--empty address in NIPS20104006
(There were 7 warnings)
```

```
bibtext _ label error
```

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

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

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

```
=====
[2017-11-06 17.35.48] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Font shape 'OMS/LinuxLibertineT-TLF/m/n' undefined using 'OMS/ntxsy/m/n' instead for sym
Some font shapes were not available, defaults substituted.
Typesetting of "report.tex" completed in 1.3s.
./README.yml
10:81    error    line too long (84 > 80 characters)  (line-length)
11:81    error    line too long (81 > 80 characters)  (line-length)
12:81    error    line too long (86 > 80 characters)  (line-length)
25:81    error    line too long (184 > 80 characters) (line-length)
25:183   error    trailing spaces  (trailing-spaces)
```

## Compliance Report

---

```
name: Ajinkya Khamkar
hid: 211
paper1: 10/27/2017 100%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
3
wc 211 paper2 3 1733 report.tex
wc 211 paper2 3 2098 report.pdf
wc 211 paper2 3 1034 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

40: The last decade has seen the rise of big data. Industries and organizations collect consumer and machinery data to make data driven business decisions. Machine learning techniques are used to drive data driven decisions in organizations. Traditional machine learning algorithms were designed prior to the advent of big data era. They are ill-equipped for handling the scale and volume of

big data tasks \cite{Papamakarios14comparisonof}. Recent advancements in hardware allow for running machine learning algorithms in parallel. In a parallel environment these algorithms suffer asynchronous gradient updates. In section \ref{data}, we discuss the need for efficient algorithms. In section \ref{traditional} we discuss traditional machine learning algorithms and their drawbacks. In section \ref{improve}, we discuss improvements to existing methods to support big data tasks. In section \ref{deploy}, we discuss various techniques to deploy these algorithms in a parallel environment for efficient and optimal performance. In section \ref{conclude}, we conclude our discussion.

42: \section{Data} \label{data}

46: \section{Traditional Algorithms} \label{traditional}

76: \section{Improvements to traditional algorithms} \label{improve}

83: Coordinate descent algorithms \cite{Wright} is a derivative-free optimization technique. to approximate the convex function  $f(x)$ , optimize one column of  $f(x)$  using  $\min_{x \in R^n} f(x)$  at every iteration The algorithm converges for strictly convex error surfaces. Algorithm \ref{cd} is the general framework of coordinate descent algorithm.

87: \caption{coordinate descent} \label{cd}

101: Cyclic coordinate descent cycles through each block and computes the descent for each block iteratively. Random Block Coordinate algorithm \cite{pmlrv37nutini15} presented in algorithm \ref{rcd} draws from a random distribution and updates the parameters.

Block Coordinate descent with Gauss-Southwell \cite{pmlrv37nutini15} rule presented in algorithm \ref{gscd} selects the block which minimizes the error in a greedy manner.

105: \caption{randomized coordinate descent} \label{rcd}

119: \caption{gauss-southwell coordinate descent} \label{gscd}

145: Instead of computing gradient per sample, splitting the dataset into multiple mini batches \cite{Bottou2010} and sampling randomly or cyclically from the mini-batches as presented in algorithm \ref{sgd} leads to much stable learning and faster convergence

149: \caption{minibatch stochastic gradient descent} \label{sgd}

166: \section{Parallel Implementation of Gradient Descent} \label{deploy}

168: Zinkevich, Weimer, Smola \& Li, 2010 \cite{NIPS20104006} introduced parallel stochastic gradient optimization technique. This technique is shown to converge and is simple to implement. Gradients generated by the workers in the network are averaged. Algorithm \ref{SGD} is applied iteratively until convergence

172: \caption{Parallel SGD ( $\{c^1, \dots, c^m\}$ , T, n,  $w_o, k$ )} \label{SGD}

```
199: \section{Conclusion} \label{conclude}

figures 0
tables 0
includegraphics 0
labels 10
refs 5
floats 0

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

Label/ref check
passed: True
```

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

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

WARNING: algorithm and below may be used improperly

99: Here  $\lambda$  represents the step size. The algorithm is simple and easily scalable. Block variants of coordinate descent algorithm are discussed below.

WARNING: algorithm and above may be used improperly

164: In the following section we discuss implementation of the above algorithms in a parallel computing environment

WARNING: algorithm and above may be used improperly

201: Machine learning algorithms play an important role for big data applications. We wished to highlight theoretical optimization

constraints for big data using traditional techniques. The sheer volume of the data leads to other optimization problems. These include feature reduction, Randomized methods for matrix decomposition over Principal Component analysis and iterative model building which are beyond the scope of this discussion. We presented the drawbacks of traditional gradient descent, which remains the backbone of machine learning algorithms. We discussed several methods to approximate the true gradient of the error function or perform gradient-free parameter updates. We also discussed several parallel implementations of the above techniques for handling big data tasks efficiently and optimally.

bibtex

---

label errors

bibtex errors

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--numpages field, but no articleno or eid field, in Kotsiantis
Warning--numpages field, but no articleno or eid field, in Meng
Warning--no number and no volume in 2012arXiv
Warning--page numbers missing in both pages and numpages fields in 2012arXiv
Warning--numpages field, but no articleno or eid field, in Wright
Warning--page numbers missing in both pages and numpages fields in DBLP
Warning--empty address in NIPS20104006
(There were 7 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 League of Legends

Junjie Lu

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

## ABSTRACT

League of Legends is the most popular MOBA online game in these years. There are millions of players all over the world. And big data about League of Legends is also deserved to be researched. The data could tell us many things we do not know. We could know usage of champions by learning weekly free champion rotation. Then getting some information of a part of income of Riot Game company. Also we could learn some advanced information of this game by researching win rate and something else. These could help players getting a better performance in the game. Furthermore, we can also try to figure out which side would win the game before it ends. All this is from big data analysis.

## KEYWORDS

I523, HID214, Big Data, League of Legends, Win rate, Income

## 1 INTRODUCTION

Computer games become more and more popular in these years. Children could play different single games on computer ten years ago. Different kinds of online games were produced in the past decade. And they have a really large market now. In 2016, market of video game all over the world is more than 111 billion.[1] MOBA game (Multiplayer Online Battle Arena) occupies the most players because of its flexibility and uncertainty. The most successful MOBA game is League of Legend. In America, it has more than 30 million players and ten times of it in worldwide. That is an amazing number meaning a large success. Player could pick a champion before the game and fight with the champion he picked as a team with five persons. There are 137 existing champions and the number is still growing. Different champions play different role in their team with their own capability. It is interesting playing different champions. Players must pay for champions so that they could use them. Riot game, the producer of League of Legends sets free 10 champions every week to give players better playing experience. If player likes these champions, they could purchase them so that they can use them anytime. Furthermore, they can also buy some skins which can give champions better appearance. These could bring Riot Game much income. We could analyze the influence of free champions on income and some other fields.

## 2 DATA ANALYSIS FOR INCOMING

We could get much data on LoLDB website. It can tell you much about champions such as win rate, pick rate and so on. Before analysis, we should know there are different tiers in League of Legends. It can tell people how intelligent a player is. We just pick from normal to Diamond. According to the data, we could get the usage of a champion just name it AB as follow:

$$U_t = 100 * \frac{M_t}{\sum_{c=1}^C \frac{W_t}{5}} \quad (1)$$

In this equation,  $M_t$  is the number of matches in which champion AB being selected in tier  $t$  on one day.  $W_t$  means total winners in tier  $t$ . [2] The usage score roughly translates to the percent of games in which a champion appears and allows for an increased score when popular champions appear on both teams.[3]

According to Figure 1, the usage of free champions is absolutely increased especially in lower tiers. And the usage of champions free on previous week is still higher than champions who are not free recently. We can get that amount of players paid for their beloved champions after using them freely.

As the data from Figure 2, we can get the probability of players who bought champions after playing with free champion rotation. And we can get:

$$I = \sum_{c=1}^{10} \sum_{t=1}^6 a * U * p_c \quad (2)$$

$I$  is the income by selling champions.  $a$  is a constant of price of champion.  $p_c$  is the probability of purchasing champions. So we can get income of selling champions of Riot Game approximately with this equation.

Furthermore, skin can also bring income for Riot Game. As a survey in China about if players are forward to purchasing a skin for champion he loves. It can not only bring a better appearance for the champions, but also some confidence when fighting with the champion. More than 30% users are willing to buy skins according to the survey in Figure 3. One champion could have different kinds of skins with different prices.

$$I_s = \sum_{c=1}^{10} \sum_{t=1}^6 b * U * p_s \quad (3)$$

$b$  is a constant for value of skins and  $p_s$  is the percentage of players buying skins. We can get  $I$  and  $I_s$  as incoming of Riot Game by selling champions and skins from data analysis.

## 3 PREDICTION FOR PATCH

All games need balance. There are more than 130 champions in League of Legends and they all have their own abilities, playing different roles. Designer should give equal ability to get some usage of players. But it is a really difficult task. When we strength champion A and its usage increases, that must mean usage of another champion decreased. Also if champion C always has a

wonderful performance when he faces champion D, the usage of champion D won't be high if usage of champion C is high. To deal with this situation, designers must fix features of champions patch next patch. They not only fix bug of the game, but also buff or debuff champions in turn to make sure they could get enough picked. In this case, designer would only adjust several champions in each patch. With more than two year observation, we found that designers prefer to adjust champions whom has higher win rate. A higher win rate means players prefer to pick them to have a easy win. This limits usage of other champions. Hence designer have to make some adjustment on the champion, such as turn down the damage it can make or increasing cool down time of skills. Designers always take this way to keep balance between champions. For example, in Patch 6.13, champion Graves has a out of power win rate 57%. This is much higher than the average and means that Graves occupies the most usage. Designers strengthened champion Kindred by increasing its armor so that it can beat Graves in late game. And increased damage of champion Nidalee encouraging it beats Graves in early games. Then we can see the win rate of these three champions are approximately equal in next patch. That is what they want. With this thought, we could use the big data of win rate and some else to predict something of patch. For instance, champion Galio has a pretty good performance in matches with a win rate of 54%. So I think champion Malzahar and Vayne will get buffed in next patch to limit win rate of Galio. There is something more what we have to consider is the impact of new champions. Most of new champions, will stand on its usage peek for one week or two, and cool down after it is free week. Then it will become as other champions. Some thing different is champion Yasuo, for unknown reason, its usage is high stably. No matter how designer weaken it, there are still millions of players loving it and fighting with it. People have to admit it is the most popular champion they have seen. What is more we have to pay attention is champion reworking. For instance, champion Sion got rework in patch 4.18 and we can see a tremendous increase in usage in October 2014. The usage stabilizes quickly overtime, and it remains mostly unaffected during his free week because many players already own Sion (due to his age) and inexpensiveness. [4]

## 4 JUDGE TOXIC BEHAVIOR

The game has its own report system, players could report others who has toxic behavior. But it is hard to say who is really toxic if two players report each other. Big data helps a lot here. The data KDA(Kill-Death-Assist) is essential in this part.

$$KDA = \frac{Kill + Assist}{Death} \quad (4)$$

Players who has toxic behavior always has a negative attitude to the match. In this case, the value of KDA is always pretty small. Number of killing minions  $cs$  could also make some contribution in this part.

$$T = \frac{KDA * cs}{R} \quad (5)$$

Number  $R$  is time of reported gotten from teammates in one game. In this case, when we got some report form players, we could use big data to judge value  $T$  to get conclusion of people who is really toxic one. The player is toxic when value of  $T$  is pretty low. System

could give toxic players some punishment and prevent from give punishment to wrong people.

## 5 PREDICT RESULT OF MATCHES FOR FURTHER WORK

These days the world championship is in China. 16 teams are fighting for the final championship. The competition is a little different from normal games. Each side could ban 5 champions in one match. And one champion could not pick twice on one side or two. There are many strategy on ban and pick, and this is coaches' job. How to pick 5 good champions after banning 10 champions totally helping winning the match. Coaches need to refer big data, not only win rate of every champion, but also win rate of champions when players in his team playing them. What is more, data from OP website could provide more data in detail such as win rate of champion A when facing champion B. Above all it needs a complex model to predict the result of a match. So it is further work.

## 6 CONCLUSION

Big data could help us to analyze much useful information. We could know how much Riot Game earn by selling champions and skins. And having a judgement to the patch. Players could practice champions would get strengthen in advance. It could help them get a higher win rate before everyone got this. And people may get result prediction of a match before it begins in the future.

## ACKNOWLEDGMENTS

The author 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] Simon Ferrari. 2013. From Generative to Conventional Play: MOBA and League of Legends.. In *DiGRA Conference*.
- [2] Yubo Kou and Bonnie A Nardi. 2014. Governance in League of Legends: A hybrid system.. In *FDG*.
- [3] Haewoon Kwak and Jeremy Blackburn. 2014. Linguistic analysis of toxic behavior in an online video game. *arXiv preprint arXiv:1410.5185* (2014).
- [4] Choong-Soo Lee and Ivan Ramler. 2015. Investigating the impact of game features and content on champion usage in league of legends. *Proceedings of the Foundation of Digital Games* (2015).

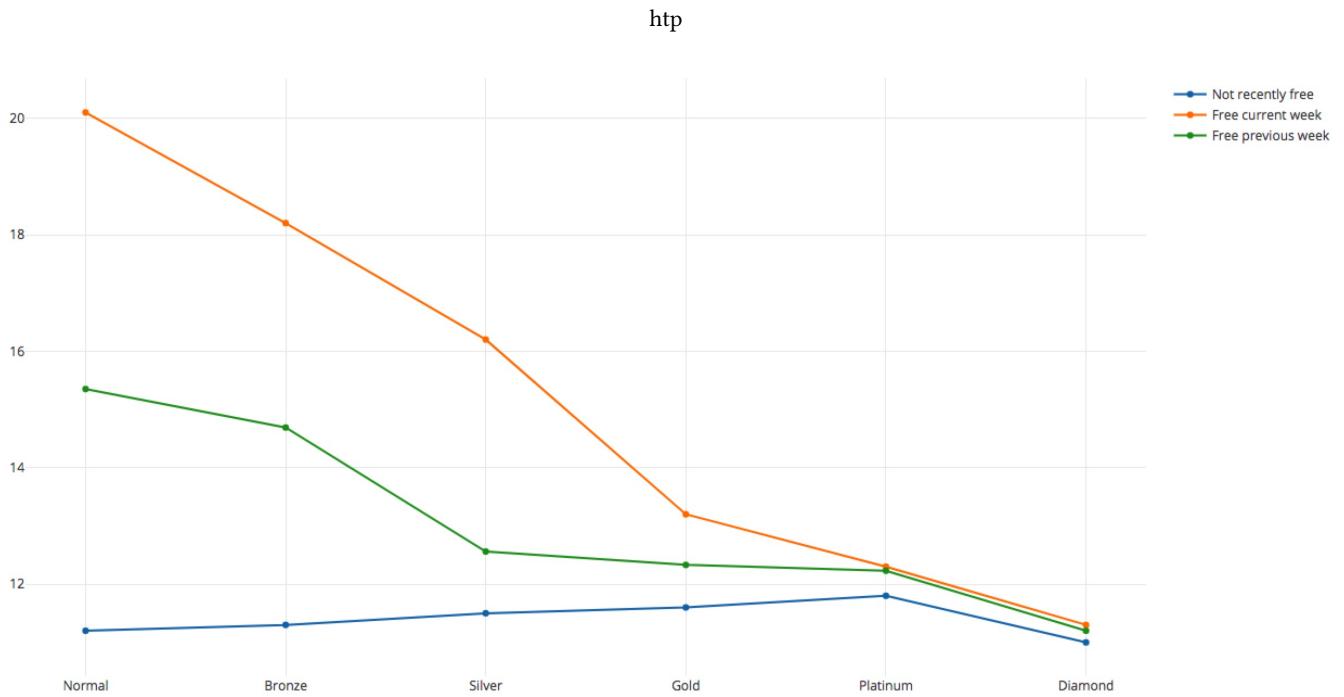
[Figure 1 about here.]

[Figure 2 about here.]

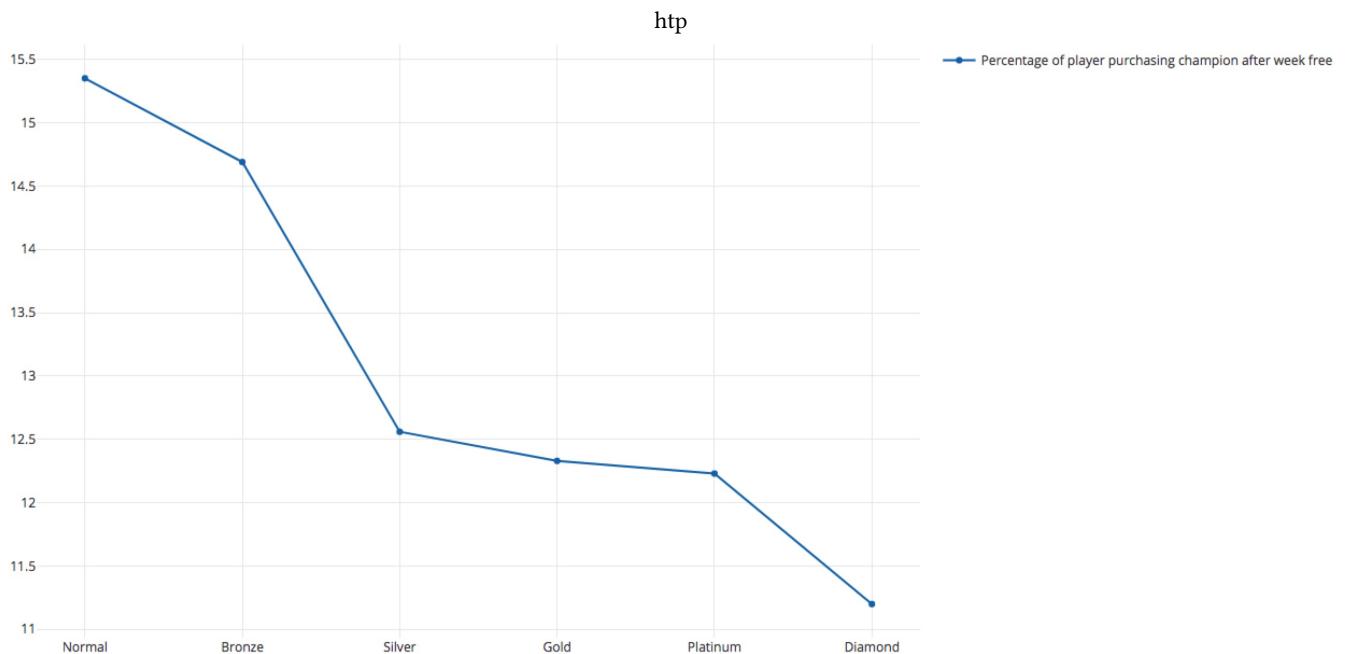
[Figure 3 about here.]

#### LIST OF FIGURES

1	Usage of free champion	4
2	Probability of purchasing champion	4
3	Percentage of players willing to buy skins	5

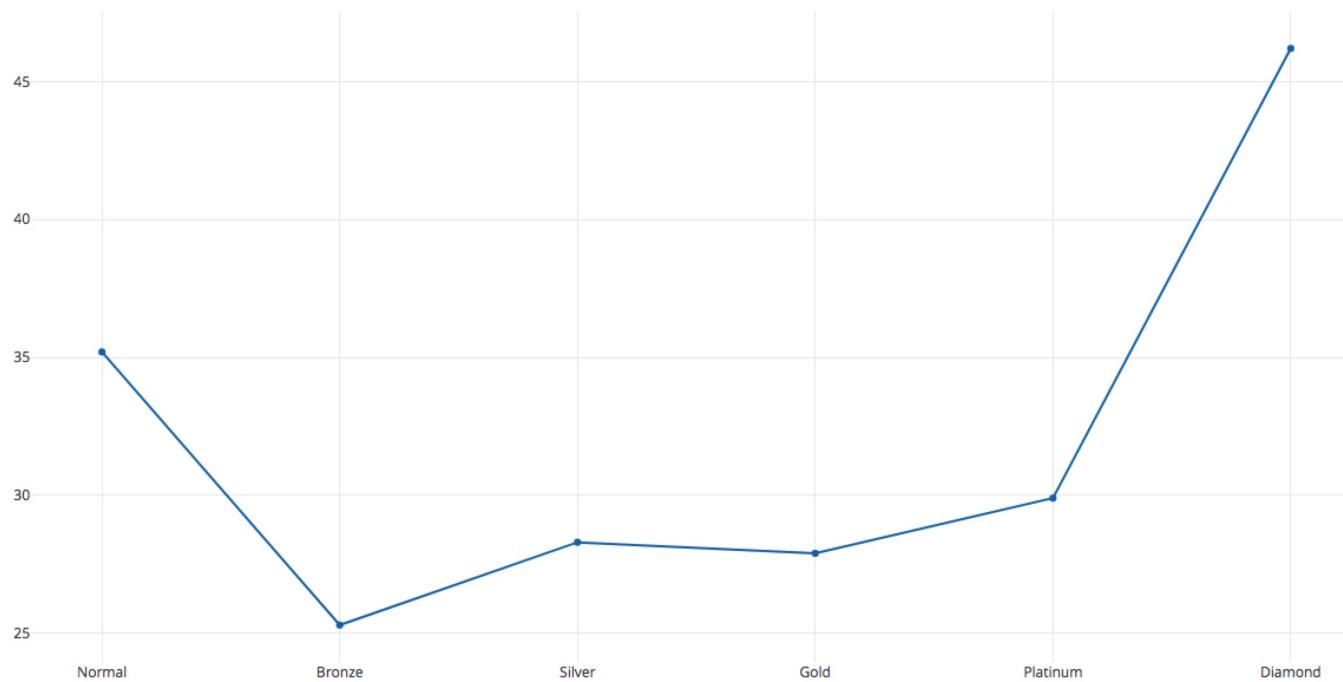


**Figure 1: Usage of free champion**



**Figure 2: Probability of purchasing champion**

htp



**Figure 3: Percentage of players willing to buy skins**

```
bibtext report
```

---

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in bbb
Warning--empty address in bbb
Warning--page numbers missing in both pages and numpages fields in bbb
Warning--empty publisher in ccc
Warning--empty address in ccc
Warning--page numbers missing in both pages and numpages fields in ccc
Warning--no number and no volume in ddd
Warning--page numbers missing in both pages and numpages fields in ddd
Warning--no number and no volume in aaa
Warning--page numbers missing in both pages and numpages fields in aaa
(There were 10 warnings)
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-11-06 17.35.54] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.0s.
```

---

```
Compliance Report
```

---

```
name: Lu, Junjie
hid: 214
paper1: 100% Oct 29th
paper2: 100%
project: 0%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
5
wc 214 paper2 5 1737 report.tex
wc 214 paper2 5 1774 report.pdf
wc 214 paper2 5 86 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
83: \begin{figure}{htp}
85: \includegraphics[width=1.0\textwidth]{images/WechatIMG112.jpeg}
87: \label{Figure 1}
89: \begin{figure}{htp}
91: \includegraphics[width=1.0\textwidth]{images/WechatIMG114.jpeg}
93: \label{Figure 2}
95: \begin{figure}{htp}
97: \includegraphics[width=1.0\textwidth]{images/WechatIMG115.jpeg}
```

```
99: \label{Figure 3}

figures 3
tables 0
includegraphics 3
labels 3
refs 0
floats 3

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

45: According to Figure 1, the usage of free champions is absolutely increased especially in lower tiers. And the usage of champions free on previous week is still higher than champions who are not free recently. We can get that amount of players payed for their beloved champions after using them freely. \\

47: As the data from Figure 2, we can get the probability of players who bought champions after playing with free champion rotation. And we can get:\\

52: Further more, skin can also bring income for Riot Game. As a survey in China about if players are forward to purchasing a skin for champion he love. It can not only bring a better appearance for the champions, but also some confidence when fighting with the champion. More than 30\% users are willing to buy skins according to the survey in Figure 3. One champion could have different kinds of skins with different prices. \\

86: \label{Figure 1}

92: \label{Figure 2}

98: \label{Figure 3}

passed: False -> labels or refs used wrong

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

find textwidth

---

85: \includegraphics[width=1.0\textwidth]{images/WechatIMG112.jpeg}

91: \includegraphics[width=1.0\textwidth]{images/WechatIMG114.jpeg}

```
97: \includegraphics[width=1.0\textwidth]{images/WechatIMG115.jpeg}
passed: False
```

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
The style file: ACM-Reference-Format.bst
Database file #1: report.bib
Warning--empty publisher in bbb
Warning--empty address in bbb
Warning--page numbers missing in both pages and numpages fields in bbb
Warning--empty publisher in ccc
Warning--empty address in ccc
Warning--page numbers missing in both pages and numpages fields in ccc
Warning--no number and no volume in ddd
Warning--page numbers missing in both pages and numpages fields in ddd
Warning--no number and no volume in aaa
Warning--page numbers missing in both pages and numpages fields in aaa
(There were 10 warnings)
```

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

```
passed: True
```

ascii

---

---

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

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

find newline

---

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

find cite {

---

passed: True

# How Big Data Transform Education

Geng Niu

School of Education Indiana University  
752 Woodbridge Drive  
Bloomington, Indiana 47408

## ABSTRACT

Educators have been searching for new approaches of teaching. In the past century, education has been tremendous progress in terms of teaching methods. However, a close look at these development reveals that development made in science and technologies drove these advances made in education. Therefore, it is very important for educators to explore the potential of big data in advancing education in this new century.

## KEYWORDS

i523, hid 218, Big Data, Education

## 1 INTRODUCTION

The development of instructional or teaching methods is closely associated with the development of educational psychology. One of the most important theory of learning is behaviorism. "Behaviorism equates learning with changes in either the form or frequency of observable performance. Learning is accomplished when a proper response is demonstrated following the presentation of a specific environmental stimulus" [6]. Behaviorists tie stimulus with behaviors. For example, when a teacher gives a student a reward, no matter it is verbal or something real, the students will study harder. In this scenario, the reward is the stimuli and studying harder is the behavior which can be observed. However, behaviorists ignore the process of learning. Cognitive theory puts more focus on how people learn. Cognitivists propose that human have sensory stores which is very limited in accepting information, and short-term memory which is reached by information after it passes sensory stores and long-term memory in which information is stored permanently so learners can retrieve it when they need it. And knowledge is categorized by procedural knowledge and declarative knowledge [2]. Instructional design, according to cognitivists, need to be made to facilitate information process and be in line with different types of knowledge. In addition, constructivism made one more step forward towards learning. Learning, according to constructivist theory, is a process of meaning making, a process of solving problems when encountering cognitive conflict and a social activity such as collaboration and negotiation [10].

## 2 A NEW AGE?

The development of education mentioned above can serve as general guidelines for educators to manage classes. However, it is not individualized. If we are adopting what could be important and impactful practices, but we really don't know, because we don't have data to inform, instrument, tune, test, and measure the impact student-level impact of our seemingly endless stream of initiatives [5]. How will big data transform education?

## 3 EDUCATION IS MORE ADAPTIVE

Students with different abilities can learn at different paces. But in a traditional classroom where students learning the same lessons by listening lectures, it is impossible to implement adaptive learning. But this will be a reality now. Institutions have access to data from various sources such as online application, classroom activity software for exercises and testing, social media, blogs and survey of staff. With the help of adaptive learning platforms, Universities can provide personalized feedback to students, monitor student satisfaction, increase attainment and give students opportunities to reflect on their own learning. Adaptive learning platforms collect and interpret data from learner interaction. And teachers will be provided with real-time reports so they can have revise their teaching strategies to ensure better outcome. In such way, educators will eliminate subjective perceptions of learners' experience and find trends in learning and teaching experience [4].

## 4 BIG DATA AND MOOCS

MOOCs stands for Massive Online Online Courses. It has become one of the most popular mode of informal learning and is considered "as an opportunity to gain access to education and professional development and to develop new skill" [7]. There are some very popular MOOCs sites one can find on the internet such as Coursera, Edx and Udemy. The online courses in these sites always have short instructional videos whose length varies from 5 to 20 minutes, and some quizzes embedded in these videos and discussion forums which may contain 2000 students. Because data in MOOCs includes longitudinal data, rich social interactions such as videoconference and detailed data about other activities, educators know have the opportunities to improve student learning in the following areas: individualized students' learning path; diagnosis of students' needs; reducing students' and institution's cost; problem-solving skills in complex context [1]. On online learning, students will generate their data trail which will be analyzed in real time so an optimal learning environment will be created. Also educators can monitor students' online activities such as how long they stay in a specific page and with such information we may know which part needs more elaboration and provide in-time support to students [8]. In addition, learning analysts can collect data about when where online learners drop a certain course to see if there is a general trend to decide what parts of the course need to be improved based on a better needs analysis.

## 5 COMPUTERIZED LEARNING

Data mining and data analytic software enable educators to get immediate feedback on how well the learners were doing online. Underlying patterns can be analyzed to foretell student outcome such as dropping out, needing extra help or being able to do more

demanding assignment. For example, a data analytic software was employed in a high school chemistry class which aimed at helping students understand the relation between submicroscopic particles and macroscopic phenomena. With the assistance of the software, teachers are able to know how students master chemistry, statistics, experimental designs, and key mathematical principles through assessment tools and pre- and post-test evaluation [9].

## 6 ADVANCING EDUCATION

People are used to be put in a certain grade according to their age. For example, in China children are typically start their first year in primary school at the age of 7. Students advance to a higher grade when they grow older. The result is that all the friends around are basically born in the same year. However, with the help of big data and data analysts, educators can find which student is learning faster and is ready to advance to a more difficult class and who need more support before he or she in a certain topic [3]. As a result, we can imagine a school where students of different ages study together in K-12 education and in undergraduate level classes.

However, such changes may have some unpredictable results. The positive result can be better school performance with exchanges of ideas from different groups and better learning effectiveness. However, some negative effects can also be predicted. Once fast learners are put together and slower learners are left in other classes, those slower learners may lose opportunity to learn from people who perform better in some subjects than them. And learning is not just to increase the input of knowledge, it also involves socializing. It is still unclear when people at different ages mix together in k-12 education whether the interaction between students will become better. Let's take China as an example. In China, the best resources are located in the eastern coast where the economy is more developed than the west. If students are categorized by their learning data, the gap of education is definitely going to be widening instead of being reduced. So the data is just helping us to make decisions and it is up to policy makers to make sure that what is better for education.

## 7 SELF-MANAGEMENT

Because of better availability of information driven by the use of mobile devices, learners today can constantly engage in informal learning. They can learn in MOOCs, read papers that they are interested, watch some tutorial videos in various websites. As a result, it is impossible for teachers or other staff at schools to monitor learners' learning process. Recommender systems will send learners courses that they may be interested at and videos that they may want to watch, which increase learning activities. So it is very necessary that some learning analysis software can help learners to review their own learning activities and even their peers learning activities in courses they take together. In this way, learners can diagnose their own learning and learn from others.

In the future, everyone will engage in mobile learning and in-time learning. People will not only learn to get certificates and get a job, but learn to solve problems popped up in their life. Moreover, they will be able to share their learning experience to others when others encounter similar problems. Therefore, a net of sharing and

learning will be created to replace the school-centered knowledge world.

## 8 MEET LEARNER'S NEEDS

For instructional designers or learning specialists, it is very important to do a thorough needs analysis before developing any instruction. However, in reality, especially in corporate learning, it is almost impossible to use survey and interviews to collect information for needs analysis. Can big data help in terms of finding employees' needs in learning something new to tackle problems at work by data mining and other means? Can big data also help us to better understand students in formal learning? We still need time to see.

## 9 CONCLUSION

Big data provide many new opportunities to improving learning in terms of extending traditional learning theories and in terms of revolutionizing education. With the help of big data, it will be easier to implement constructivist theory in learning, and help analyze learning in ways which cannot be done in the past. However, we should also note that with opportunities comes some potential threats such as widening the disparities between the well-learned and the ill-learned.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Chris Dede and Andrew Ho. 2016. Big Data Analysis in Higher Education: Promises and Pitfalls. *Educause Review* 51, 5 (2016). <https://er.educause.edu/articles/2016/8/big-data-analysis-in-higher-education-promises-and-pitfalls>
- [2] WELLESLEY R. Foshay Kenneth H. Silber. 2006. *Handbook of human performance technology: principles, practices, and potential*. Wiley, Chapter 16, 370–413.
- [3] Dan Kerns. 2013. 10 Ways Big Data is Changing K-12 Education. (2013). <http://www.dreambox.com/blog/10-ways-big-data-changing-k-12-education-2>
- [4] How Big Data Will Boost Learning and Teaching in Higher Education. 2016. Cogbooks. (2016). <https://www.cogbooks.com/2016/10/05/big-data-will-boost-learning-teaching-higher-education/>
- [5] Mark D. Milliron. 2016. Higher Education's Turn To Big Data For 'Healthy' Change. (2016). <https://www.forbes.com/sites/schoolboard/2016/09/16/higher-educations-turn-to-big-data-for-healthy-change/#20170a81bcc>
- [6] Newby Peggy A. Ertmer, Timothy J. 1993. Behaviorism, cognitivism, constructivism: comparing critical features from an instructional design perspective. *Performance improvement quarterly* 6, 50-72 (1993), 50.
- [7] Stephanie D. Teasley Tawanna Dillahunt, Zengguang Wang. 2014. Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *International Review of Research in Open and Distance Learning* 15, 5 (2014), 20.
- [8] Mark van Rijmenam. 2016. Four Ways Big Data Will Revolutionize Education. (2016). <https://datafloq.com/read/big-data-will-revolutionize-learning/206>
- [9] Darrell M. West. 2012. Big Data for Education: Data Mining, Data Analytics, and Web Dashboards. (2012), 2.
- [10] Brent G. Wilson. 2012. *Constructivism in practice and historical context*. Pearson, Chapter 5, 45.

## 10 ISSUES

DONE:

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

## 10.1 Assignment Submission Issues

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

## 10.2 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

## 10.3 Formatting

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

Other formatting issues

## 10.4 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}

## 10.5 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

## 10.6 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

## 10.7 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 inflate your paper if you are below the page limit

## 10.8 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 float. 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

```
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--page numbers missing in both pages and numpages fields in Dede2016
Warning--can't use both author and editor fields in KennethH.Silber2006
Warning--empty address in KennethH.Silber2006
Warning--no journal in Kerns2013
Warning--no number and no volume in Kerns2013
Warning--page numbers missing in both pages and numpages fields in Kerns2013
Warning--no journal in Learning2016
Warning--no number and no volume in Learning2016
Warning--page numbers missing in both pages and numpages fields in Learning2016
Warning--no journal in Milliron2016
Warning--no number and no volume in Milliron2016
Warning--page numbers missing in both pages and numpages fields in Milliron2016
Warning--no journal in Rijmenam2016
Warning--no number and no volume in Rijmenam2016
Warning--page numbers missing in both pages and numpages fields in Rijmenam2016
Warning--no journal in West2012
Warning--no number and no volume in West2012
Warning--empty address in Wilson2012
(There were 18 warnings)
```

```
bibtext _ label error
```

---

```
bibtext space label error
```

---

```
bibtext comma label error
```

---

```
latex report
```

---

```
[2017-11-06 17.36.00] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
```

## Compliance Report

name: Niu, Geng  
hid: 218  
paper1: 100%  
paper2: 100%

## yamlcheck

```
wordcount
```

```
4
```

```
wc 218 paper2 4 1601 report.tex  
wc 218 paper2 4 2387 report.pdf  
wc 218 paper2 4 314 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
```

```
passed: True
```

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

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

```
passed: True
```

```
floats
```

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

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

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

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

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)  
The top-level auxiliary file: report.aux  
The style file: ACM-Reference-Format.bst  
Database file #1: report.bib  
Warning--page numbers missing in both pages and numpages fields in Dede2016  
Warning--can't use both author and editor fields in KennethH.Silber2006  
Warning--empty address in KennethH.Silber2006  
Warning--no journal in Kerns2013  
Warning--no number and no volume in Kerns2013  
Warning--page numbers missing in both pages and numpages fields in Kerns2013  
Warning--no journal in Learning2016  
Warning--no number and no volume in Learning2016  
Warning--page numbers missing in both pages and numpages fields in Learning2016  
Warning--no journal in Milliron2016  
Warning--no number and no volume in Milliron2016  
Warning--page numbers missing in both pages and numpages fields in Milliron2016  
Warning--no journal in Rijmenam2016  
Warning--no number and no volume in Rijmenam2016  
Warning--page numbers missing in both pages and numpages fields in Rijmenam2016  
Warning--no journal in West2012  
Warning--no number and no volume in West2012  
Warning--empty address in Wilson2012
```

(There were 18 warnings)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

non ascii found 8220  
non ascii found 8217  
non ascii found 8217  
non ascii found 8221  
non ascii found 8217  
non ascii found 8217

---

The following tests are optional

---

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Applications in Aviation Industry

Swargam, Prashanth  
Indiana University Bloomington  
107 S Indiana Ave  
Bloomington, Indiana 47408  
pswargam@iu.edu

## ABSTRACT

Data generated by aviation industry is being increased enormously. The data generated by all the components of aviation industry can be analysed for reducing the operational costs, predict customer behaviour, analyse customer satisfaction. These applications of big data in aviation industry makes it a prominent player. Hence, collecting this data, storing and processing them for desired results can help the aviation industry in boosting their profits and improve customer satisfaction. Various applications of Big data, their challenges and models are discussed here.

## KEYWORDS

HID228, I523, Big Data, Aviation Industry, Analytics ,

## 1 INTRODUCTION

Big Data has transformed the businesses were being conducted. Every sector is integrated with data and generating huge amounts of data every day. All the companies are following data driven approach to crunch their competition. With the advent of concepts like Internet of things, the generation of data is increasing by many folds. This brings scope for a new business which handles the storage and analysis of data.

The solutions offered by Big Data in many industrial sections have revolutionised the respective businesses. In aviation industry, where data as big as 20tb is generated from an aircraft flying for an hour[? ]. Big Data can offer influential solutions in terms of dealing with the massive data. This data ,if is processed in an efficient way would increase the customer satisfaction at a reduced running and operational costs which in turn increases the profits.

## 2 APPLICATIONS

### 2.1 Baggage Handling

All the customer check-in their bag and have a doubt if their bags are being transported with them. There are several cases where customers raise some complaints about their bag being missed or bag transported to another destination. Traditional barcode system was used to handle this task. As the number of airline users increased, this solution was not profitable for customers and airline operators. However, this is being replaced by the new technology which uses radio frequencies to track real-time location of the bag. Bags which are checked in at the kiosk are assigned with a microchip. These chips will send the data related to the location of the bag frequently. The data generated by these chips is processed and stored. The processed data is available to the customers through mobile application or a web interface[? ].

### 2.2 Flight Safety

All the flights have many sensors which generates a lot of data related to flight status and incidents. According to, a Being 737 generates nearly 20tb of data for one hour and an average cross international plane travelling for 6 hours generates 240 tb of data. Most of these data is related to safety and status of various equipment on the flights. A lot of this data should be filtered and mined to generate a meaningful and usable data. Southwest Airlines partnered with NASA for crunching this data and generating a meaningful data. NASA uses machine learning algorithms to mine this data [? ].

This collected data from the flight can be analysed to decide a desired value for variables like altitude, wind speed, thrust, weight of the aircraft are proposed to the pilot for increased fuel economy. This data can also be helpful in deciding the nature of services according to the nature of the location and fuel costs [? ].

### 2.3 Personalized promotion

In the advent of smart devices, all the industries including airline industry have come closer to the customer. Variables which are considered as characteristics are studied from the customer data available through their interaction with customers. These details range from preferences to behaviour of the customer. This data is analysed to study the behaviour of the customer and improve his experience with the airline industry [? ].

### 2.4 Pricing strategies

Pricing is an important strategy to generate profits. It is quite often to see a price bump of the airfare during the payment or checkout process. This is because of increase in demand for the journey. This demand data is analysed in the servers and a revised price is shown on the customers screen in less than minute. This calculations and analysis requires high computing power and efficient algorithms. EasyJet has uses artificial intelligence to determine the price of seat based on demand [? ].

## 3 DATA SOURCES

### 3.1 In-Flight Data

QAR Data: Quick Access Recorder records the statistics of the flight like speed, height, speed, altitude, at any instance during flight [? ]. This data is stored in servers and processed

ACARS Data: Aircraft Communications Addressing and Reporting System is a online data transmission system which transmits data to ground through the aircraft's satellite communication system[? ]. ACARS records values of different parameters during a event. An event is an action performed by the aircraft. The sensors mounted on the flights' brakes, wings, doors, send data to

the ground staff using ACARS. Aircraft connection sensors and equipment monitoring system also uses this ACARS to transmit the data to ground.

### 3.2 Data from mobile and web applications

Now-a-days all the customer interactions with airline industry is through web. All the web applications and mobile applications which are developed for interacting with customer are smart enough to store the variables which are used to study the customer behaviour [? ]. This portion of data sources generates the data at increasing rates due to the evolution of customer interaction with internet.

### 3.3 Historical Data

Data available from the previous analytics and recordings constitutes a major portion. These are generally excel sheets or other forms of data stored in servers or files. These can be used for predictive analysis of the flight.

### 3.4 Other Sources

Other sources like weather sensors, internet, analysis from third party vendors which help airline industry in scheduling and predicting flight delays .

## 4 CLOUD STORAGE IN AVIATION INDUSTRY

According to[? ], Cloud computing is an IT process in which a specific application which belong to an organisation or individual is hosted on shared pool of servers. This data storage in these shared pools of servers can be provisioned on demand and can be resized according to the change in needs.

As aviation is industry produces enormous amounts of data, this requires high capacity computing servers to store and access them on demand. This makes the local storage and maintenance costly and time consuming. Airlines can outsource this activity by hosting their applications on cloud storage on any of the vendors either on public or private clouds[? ]. This model helps them in reducing costs and heavy IT infrastructure maintenance and allows more room for them to concentrate on their own business.

In aviation industry, data is very rapid and requires faster storage and retrieval of data for efficient transactions. As these are shared pool of servers, these servers experience heavy data traffic. These servers are equipped with efficient load balancers. This ensures high availability and high speed of data access.

Customer Service can be increased without the increase in the IT infrastructure or IT workforce. As the infrastructure is outsourced, there will be minimal downtimes for activities like upgrades and maintenance This improves the customer experience[? ].

Aircraft maintenance involves changing and repairing various components of aircraft while keeping a complete track of all these changes and replacements. Cloud computing technologies offer good solutions and reduces the complexity of maintenance[? ]. Architectures are developed to track these changes and providing this information and analytics to the appropriate people on their devices.

Companies like Virgin Atlantic, Endeavour Air has implemented these solutions to enhance their Aircraft maintenance[? ].

## 5 CHALLENGES IN IMPLEMENTING BIG DATA

### 5.1 Data Size

With the advent of Internet of things, all the components of the aircraft are getting connected to the Internet. This enables communication between with the airline components and the ground staff. Every component detects its status and communicates it to the respective department. This results in generation of data and complexities in handling this huge data. According to [? ], a Virgin Atlanticfis Boeing 787 generates approximately half terabyte of data per week.

### 5.2 Data Format

Data is produced in various formats by their respective sources. This results in high complexity in calculation and visualization. This complexity increases when, data needs to be transferred among various data sources or physical cloud databases[? ].

### 5.3 Security

In Aviation Industry, a large amounts of customer data are stored and processed for better analytics in marketing and promotion, this involves a huge risk. Almost a hundred and fifty parameters related to customer is taken from their interactions[? ], this data can be misused to locate customer by the attacking agents

As huge amounts of data related to aircraft and its maintenance is analysed and stored, this data can be vulnerable and can hamper aircraft security.

Advent of Cloud computing not only brings in advantages like lower IT maintenance cost, ease of other maintenance activities, but also brings in security issues alongside. Security measures higher than Industry standards must be used to protect data in shared pools of servers.

## 6 CONCLUSION

Big Data Analytics has provided impactful solutions in computing, storing and managing large amounts of data while lowering infrastructure costs and maintenance costs. This brief summary provides a overview of application of Big Data technologies in various aspects of Aviation Industry like Baggage handling, Aircraft safety, Customer Experience and Marketing. The discussion expands to various kinds to data sources for aviation industry and gives information about advantages of cloud storage. Though Big Data Analytics have provided prominent solutions in Aviation industry, it still has challenges related to security and protection of data. Research is being conducted in these areas to make more secure.

```
bibtext report
```

```
=====
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
I found no \bibstyle command---while reading file report.aux
(There was 1 error message)
make[2]: *** [bibtex] Error 2
```

```
latex report
```

```
=====
[2017-11-06 17.36.10] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
p.1  L46   : [Maire2017] undefined
p.1  L52   : [Miller2017] undefined
p.1  L56   : [Smalley2012] undefined
p.1  L59   : [7889557] undefined
p.1  L63   : [EXASTAX2017] undefined
p.1  L67   : [EXASTAX2017] undefined
p.1  L74   : [7015483] undefined
p.1  L76   : [Wikipedia] undefined
p.2  L80   : [EXASTAX2017] undefined
p.2  L90   : [Wikipedia2017] undefined
p.2  L92   : [6548579] undefined
p.2  L96   : [Ferkoun2015] undefined
p.2  L98   : [Ferkoun2015] undefined
p.2  L100  : [Vault2014] undefined
p.2  L110  : [Finnegan2013] undefined
p.2  L114  : [6548579] undefined
p.2  L116  : [EXASTAX2017] undefined
Missing character: ""
There were undefined citations.
Typesetting of "report.tex" completed in 0.8s.
./README.yml
33:75      error    trailing spaces  (trailing-spaces)
36:81      error    line too long (82 > 80 characters)  (line-length)
36:82      error    trailing spaces  (trailing-spaces)
```

```
40:81      error      line too long (89 > 80 characters)  (line-length)
```

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

```
name: Swargam, Prashanth  
hid: 228  
paper1: Oct 20 17 100%  
paper2: Nov 06 17 100%
```

```
yamlcheck
```

```
wordcount
```

```
(null)  
wc 228 paper2 (null) 1637 report.tex  
wc 228 paper2 (null) 1541 report.pdf  
wc 228 paper2 (null) 520 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
```

```
13: \renewcommand\footnotetextcopyrightpermission[1]{} % removes  
      footnote with conference information in first column
```

```
passed: False
```

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

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

```
passed: True
```

```
floats
```

---

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

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

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

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

---

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

---

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
The top-level auxiliary file: report.aux
I found no \bibstyle command---while reading file report.aux
(There was 1 error message)
```

```
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
```

---

```
=====  
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 is 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 wash 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 wash 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 obtain 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 initiative lies 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.org/jp/ictc/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](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 inflate 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 float. 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

domestic  
reuse

irrigation

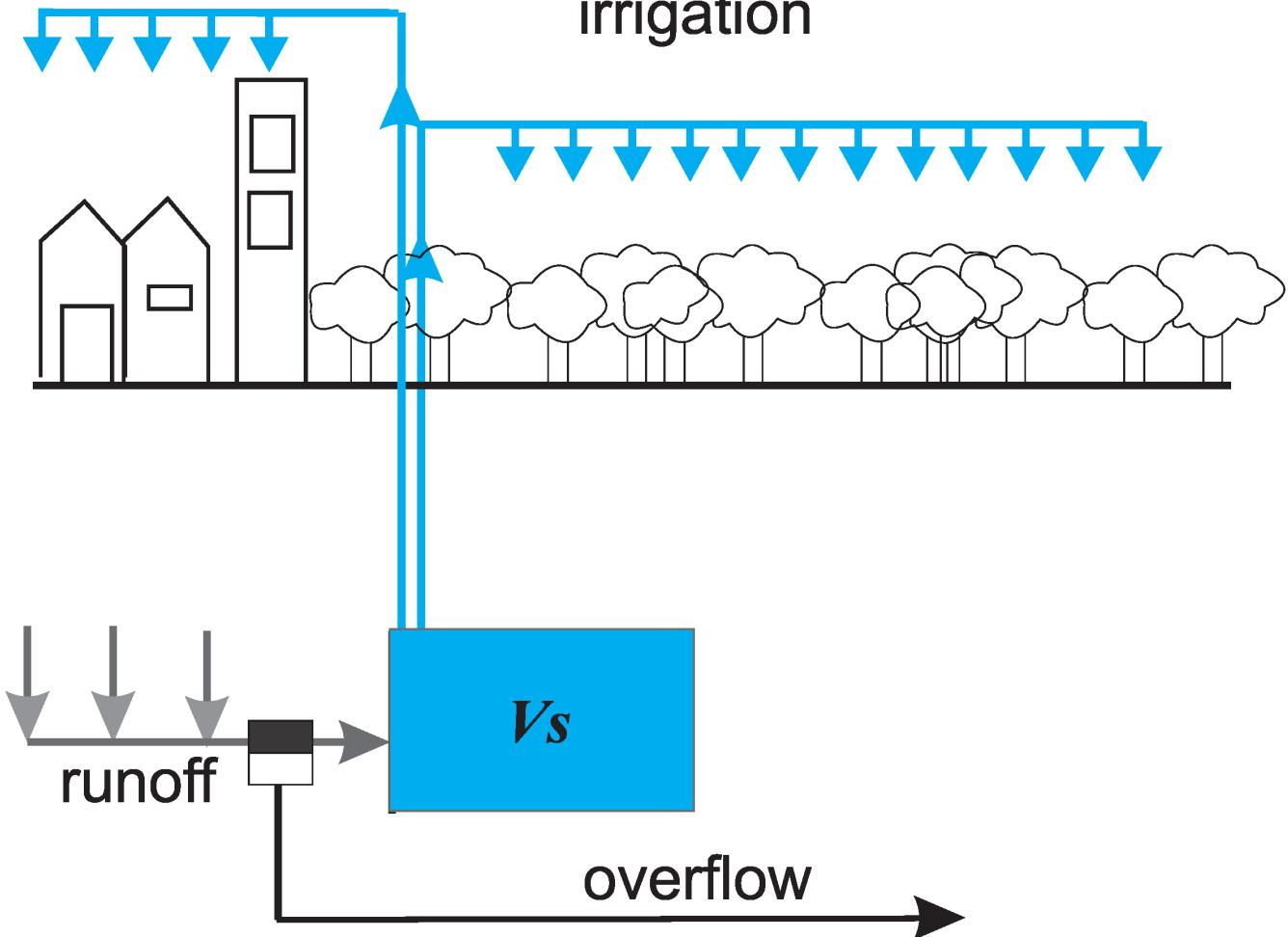


Figure 1: rainwater harvesting figure

## bibtex report

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

bibtext \_ label error

bibtext space label error

bibtext comma label error

# latex report

## Compliance Report

```
name: Rahul Velayutham  
hid: 232  
paper1: 2017-10-29 100%  
paper2: 100%  
project: in progress
```

yamlcheck

---

wordcount

---

5

wc 232 paper2 5 1768 report.tex  
wc 232 paper2 5 2448 report.pdf  
wc 232 paper2 5 151 report.bib

find "

---

passed: True

find footnote

---

passed: True

find input{format/i523}

---

4: \input{format/i523}

passed: True

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 referred to: (refs >= labels)
```

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

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

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

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

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

passed: True

ascii

---

---

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

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

find newline

---

passed: True

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

find cite {

---

passed: True

# **Big Data and Cloud Computing in Health Informatics for People with Disabilities**

Weixuan Wang

Indiana University Bloomington

Bloomington, Indiana 47405

wangweix@indiana.edu

## **ABSTRACT**

This is my abstract.

## **KEYWORDS**

Big Data, Cloud Computing, Disability informatics, Health informatics, i523, HID234

## **1 INTRODUCTION**

People with disabilities are a group that has been overlooked for a long time. Some might question that why we should care about people with disability. According to UTHHealth, as of February 2015, there are about a billion people with disabilities in the world and in the United States alone, there are 56.7 million people with disabilities [3]. Notably, the number of people with disabilities is expected to increase as a result of increasing life-span, decreases in communicable diseases, improved medical technology, and improved child mortality [7]. According to United Nation, people with disabilities are the largest minority group in the world [1–3]. While some forms of disabilities might be genetic, but temporary or permanent disabilities can happen to anyone, such as spinal cord injury after car accident, or limited mobility at later stage of life [3]. Therefore, improving the living condition and quality of life for people with disabilities are extremely important to everyone.

There are many different definition of disabilities from different organizations. The most cited official definition is the 1976 definition of the World Health Organization [1]: “An impairment is any loss or abnormality of psychological, physiological or anatomical structure or function; a disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being; a handicap is a disadvantage for a given individual, resulting from an impairment or a disability, that prevents the fulfillment of a role that is considered normal (depending on age, sex and social and cultural factors) for that individual”. While people with disabilities are those people who have limitations in their actions or activities resulting from physical or mental impairments, however, there are many types and levels of disabilities and their actions and activities are affected differently by their disabilities [1]. This presents difficulties and challenges to accommodate the different needs of people with disabilities and improve their qualities of life [5].

The development of digital technology has changed many people's lives, the life of people with disabilities has also been improved by technology. People with poor visions can using cell phones to contact others, access information online with screen readers. People with hearing problems can text other people with cell phone. This study is trying to help people with disabilities from a health informatics prospective, and looking into how big data and edge

computing can help monitor and evaluate and understand people with disabilities and help improve their quality of life.

## **2 TYPES OF DISABILITY**

Disability has different function types and levels of degrees, while these types are not completely exclusive, most of functional types of disability can be categorized into three groups: mobility, sensory, and cognitive [1]. This section provides a simple overview of these three functional types of disabilities and its challenges for people with disabilities.

Mobility problems are faced by people with physical motor disabilities (such as spinal cord injury after traumatic injury), and people have impaired muscle controls [7]. These people might have problem using technologies than are design to assist them such as wheelchairs or computer interface aids [1].

Sensory disability includes both visual and aural impairment [1]. Their conditions can range from correctable (such as using eyeglass or hearing aids) to not correctable (such as blind or deaf) [6]. Braille has been traditionally used by people with visual impairment, but now was replaced by technology such as voice synthesis and screen readers [6].

Cognitive disabilities in general refer to people with cognitive impairment who have difficulties than an average individual with one or more types of mental tasks involving language, memory, perception, problem-solving, hand-eye coordination, conceptualizing, attention and executive functions[1].

## **3 DISABILITY INFORMATICS**

Disability informatics is a sub-specialty health informatics that is defined as any application that collects, manages, and distributes information that are related to people with disabilities, as well as to care givers (including familiar members and health care providers) and rehabilitation professionals [1]. Disability informatics is closely related to medical informatics, public health informatics and consumer health informatics, because people with disabilities usually have some secondary medical condition such as poor health status and increased personal health care needs. Gather medical and health information can help to better understand and accommodate people with disabilities. A previous research has designed and deployed an extended version of Artemis system (a cloud system designed to acquire data and store physiological data of clinical information for real-time analytics) in a hospital. They have identified that high speed physiological data produced at intensive care units as big data, and the proper use of such data can promote health, reduce mortality and disability rates of critical condition patients and create new cloud-based health analytics [4]. Research also has

shown that many disabilities are genetic, therefore, bioinformatics has implications in the education of genetic screening and gene therapy treatments in the future [1].

The two main components in disability informatics are assistive technology and information technology.

The information technology applications and issues in disability informatics are categorized into four areas: virtual, personal, physical and social/intellectual.

## 4 CONCLUSION

Put here an conclusion. Conclusions and abstracts must not have any citations in the section.

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Richard Appleyard. 2005. *Disability Informatics*. Springer New York, New York, NY, Chapter chapter 11, 129–142. [https://doi.org/10.1007/0-387-27652-1\\_11](https://doi.org/10.1007/0-387-27652-1_11)
- [2] Simon Darcy. 2010. Inherent complexity: Disability, accessible tourism and accommodation information preferences. *Tourism Management* 31, 6 (2010), 816 – 826. <https://doi.org/10.1016/j.tourman.2009.08.010>
- [3] Lex Frieden. 2015. Why Disability Informatics? (02 2015). <https://sbmi.uth.edu/blog/feb-15/021115.htm>
- [4] H. Khazaei, C. McGregor, M. Eklund, K. El-Khatib, and A. Thommandram. 2014. Toward a Big Data Healthcare Analytics System: A Mathematical Modeling Perspective. In *2014 IEEE World Congress on Services*. IEEE, Anchorage, AK, USA, 208–215. <https://doi.org/10.1109/SERVICES.2014.45>
- [5] Gabriel Pestre. 2016. Big Data and Disability, Part 1. Data Pop Alliance. (March 2016). <http://datapopalliance.org/big-data-and-disability-part-1/> Accessed 2017.
- [6] Paraskewi Riga and Georgios Kouroupetroglou. 2013. Indoor Navigation and Location-Based Services for Persons with Motor Limitations. In *Disability Informatics and Web Accessibility for Motor Limitations*. IGI Global, Greece, 202–233. <https://doi.org/10.4018/978-1-4666-4442-7.ch006>
- [7] Ralph W. Smith. 1987. Leisure of disable tourists: Barriers to participation. *Annals of Tourism Research* 14, 3 (1987), 376 – 389. [https://doi.org/10.1016/0160-7383\(87\)90109-5](https://doi.org/10.1016/0160-7383(87)90109-5)

```
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
```

```
[2017-11-06 17:36:38] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Typesetting of "report.tex" completed in 0.9s.
./README.yml
9:81      error    line too long (86 > 80 characters)  (line-length)
23:81     error    line too long (135 > 80 characters)  (line-length)
28:27     error    trailing spaces  (trailing-spaces)
29:81     error    line too long (91 > 80 characters)  (line-length)
34:81     error    line too long (82 > 80 characters)  (line-length)
34:82     error    trailing spaces  (trailing-spaces)
35:81     error    line too long (82 > 80 characters)  (line-length)
35:82     error    trailing spaces  (trailing-spaces)
36:81     error    line too long (87 > 80 characters)  (line-length)
37:81     error    line too long (86 > 80 characters)  (line-length)
37:86     error    trailing spaces  (trailing-spaces)
38:81     error    line too long (84 > 80 characters)  (line-length)
38:84     error    trailing spaces  (trailing-spaces)
39:81     error    line too long (86 > 80 characters)  (line-length)
39:86     error    trailing spaces  (trailing-spaces)
41:1      error    trailing spaces  (trailing-spaces)
47:12    error    too many spaces after colon  (colons)
47:81    error    line too long (108 > 80 characters)  (line-length)
```

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

```
name: Weixuan Wang
hid: 234
paper1: Oct 22 2017 100%
paper2: Nov 8 2017 70%
project: 1%
```

```
yamlcheck
-----
```

```
wordcount
-----
```

```
2
wc 234 paper2 2 940 report.tex
wc 234 paper2 2 1029 report.pdf
wc 234 paper2 2 1427 report.bib
```

```
find "
```

```
-----
```

```
passed: True
```

```
find footnote
-----
```

```
passed: True
```

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

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

```
passed: True
```

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

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

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

=====

The following tests are optional

=====

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big data: An Opportunity for Historians

Yujie Wu

Indiana University Bloomington

Bloomington, Indiana 47401

yujiwu@iu.edu

## ABSTRACT

In human history, catastrophe, wars, big event led to a incredible loss of life. Historians collected the data of life but ignored to do the statistical analysis due to their non-statistical background. In this new age, Big data provides an insight for historians to learn humanity based on the data from the history. This paper involves the brief introduction of historical events, big data analysis based on the historical information, and the results from the big data learning.

## KEYWORDS

i523, HID235, history, Titanic, ID3

## 1 INTRODUCTION

With the rapid development of information technology, historians now enter the age of big data. Big data refers to a tremendous amount of information that produced by human and nonhuman activities in the past. The scale of big data is large enough that it is impossible for individuals to collect and preprocess the data. Hence, history which is derived from human engagement with the past must have some affinity with big data and the computer technology it represents[1]. Such instinctive property of big data provides an absolutely new perspective for historians to study and re-evaluate the history.

In this age of explosion of information, zillions of pieces information is stored on the Internet. The volume, velocity, variety, value, veracity of data are the treasure for historian to mine. But in most of time, data is neither straightforward substance nor transparent material for historians to squeeze the interesting information since they are not well-organized into a meaningful format that let the algorithm analyze them for answering the questions that historians are interested in[1]. Processing meaningless data into a coherent argument is not an easy job. Furthermore, using proper infrastructure and algorithm is difficult as well. As a historian, big data is an opportunity but also a big obstacle for the future researches.

## 2 PREPROCESS DATA

Preprocessing data is a big challenge for historians. Here is an example to illustrate a proper approach to preprocess history-relative data. The data to be introduced in this paper is from the world-famous tragedy – Titanic. Titanic was one of three “Olympic Class” liners which were an incredible feat

of engineering and ambition in their age. Titanic was the largest, fastest, and most luxurious liner. Its maiden voyage was from Southampton to New York with a lot of people on board including millionaires, movie stars, teachers and labors who were looking for a better life in United States. However, it struck an iceberg and sank in Atlantic Ocean five days after the beginning of its journey. The collision tore a series of holes along side of the hull. The sea water came into Titanic and less than three hours later, Titanic sank down about 2 miles to the bottom of the Atlantic ocean. Overall 1502 out of 2224 on broad passengers and crew lost their lives in this shocking tragedy[2].

This tragedy attracted the attention of international community. People were wondering the reasons that how such technique marvel encountered this tragedy. One of the well-known reasons that the sinking of Titanic led to such loss of life was that there were not enough lifeboats for the passengers and crew[3]. Another reason was that on the night of Sunday April 14 1912, the Atlantic ocean was flat calm, the sky clear and moonless, and the temperature was freezing-cold[2]. The weather condition was very difficult for captain and other crew to detect an iceberg. Therefore, such weather condition explained the reason why the alarm of iceberg in front was made only 40 seconds before Titanic crashed the iceberg. It was impossible for such big mechanical monster to provide a stop response. Unfortunately, Titanic accelerated towards to iceberg directly and tore a series of large holes along side of the hull.

The information from the passengers and crew on board was collected later on for historians to study one more interesting field that what sorts of people were possibly to survive.

The description of the data used to study for historians in this example is as follows. The data set has 12 attributes (columns) shown in the following table[3].

Variable	Definition
survival	Survival
pclass	Ticket class
PassengerId	ID of each passenger
sex	gender
Age	Age in years
ticket	Ticket number
sibsp	number of siblings / spouses aboard the Titanic

Variable	Definition
parch	number of parents / children aboard the Titanic
name	name of each passenger
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

Survival attribute for this example will be the label for classification. It has two values 0 for not survived and 1 for survived. Pclass attribute has 3 possible values 1 for upper class, 2 for middle class, and 3 for lower class. Age attribute is fractional if less than 1. If the age of a passenger is estimated, is it in the form of “xx.5”. Sibling defined in this data set is brother, sister, stepbrother, and stepsister. Spouse defined in this data set is husband and wife. Parent defined in this data set contains mother and father. Child in this data set includes daughter, son, stepdaughter, and stepson[3].

After defining the data value and storing the data in a algorithm-readable format, historians should process the missing values and noise which is the most significant step before analyzing or mining the data by an algorithm. Noise usually refers to non-systematic error. Such error is not caused by the algorithm or the classifier system. It is from the training dataset. For example, two tuples has the identical values in all attributes, but their label is different. It causes the inadequate attributes. To deal with the noise, the historians could delete such tuples.

If there is missing value in an attribute, the mean value is usually used as the substitution. However there is a more technological approach to fill in an unknown value by using the information provided by context. For example, the historians could use a Bayesian formalism to figure out the probability of a possible value say  $A_i$  in attribute A. In addition, decision tree approach is another way to determine the missing value. Assume  $C_s$  is a subset of C consisting of an attribute and the label. the historians could construct a decision tree based on this subset and predict the missing value using the tree model.

### 3 ANALYSIS AND IMPLEMENTATION OF ID3 ALGORITHM

The label (survival attribute) of the data set is binary value. All the attributes contain discretely numerical value. Therefore, ID3 algorithm is the best algorithm to be employed for mining and analyzing the data set. ID3 algorithm is well known as Iterative Dichotomiser 3 algorithm invented by Ross Quinlan.

### 4 CONCLUSION

This paper ...

### REFERENCES

- [1] James Grossman. 2012. "Big Data": An Opportunity for Historians? Online. (3 2012). <https://www.historians.org/publications-and-directories/perspectives-on-history/march-2012/big-data-an-opportunity-for-historians>
- [2] BBC history. 2017. The Rise and Fall of Titanic. Online. (11 2017). <http://www.bbc.co.uk/history/titanic>
- [3] Kaggle. 2015. Titanic: Machine Learning from Disaster. Online. (11 2015). <https://www.kaggle.com/c/titanic>

```
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
```

```
=====
[2017-11-06 17:36:43] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Typesetting of "report.tex" completed in 0.8s.
./README.yml
9:81      error    line too long (86 > 80 characters)  (line-length)
19:1       error    trailing spaces  (trailing-spaces)
23:81      error    line too long (89 > 80 characters)  (line-length)
23:89      error    trailing spaces  (trailing-spaces)
24:77      error    trailing spaces  (trailing-spaces)
25:81      error    line too long (106 > 80 characters) (line-length)
25:106     error    trailing spaces  (trailing-spaces)
26:81      error    line too long (109 > 80 characters) (line-length)
26:109     error    trailing spaces  (trailing-spaces)
34:9       error    trailing spaces  (trailing-spaces)
35:1       error    trailing spaces  (trailing-spaces)
38:81      error    line too long (88 > 80 characters) (line-length)
38:88      error    trailing spaces  (trailing-spaces)
39:81      error    line too long (87 > 80 characters)  (line-length)
39:87      error    trailing spaces  (trailing-spaces)
40:49      error    trailing spaces  (trailing-spaces)
47:9       error    trailing spaces  (trailing-spaces)
```

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

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

```
yamlcheck
-----
```

```
wordcount
-----
```

```
2
wc 235 paper2 2 1221 report.tex
wc 235 paper2 2 1099 report.pdf
wc 235 paper2 2 76 report.bib
```

```
find "
```

```
passed: True
```

```
find footnote
-----
```

```
12: \renewcommand\footnotetextcopyrightpermission[1]{} % removes
      footnote with conference information in first column
```

```
passed: False
```

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

```
passed: False
```

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

```
figures 0
```

```
tables 0
includegraphics 0
labels 0
refs 0
floats 0

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

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

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Analytics for Social Media Threat Intelligence

Tousif Ahmed  
Indiana University  
150 S Woodlawn Avenue  
Bloomington, Indiana 47405  
touahmed@indiana.edu

## ABSTRACT

Social media has become a virtual society for everyone where billions of people are interacting every day. With the humongous number of users, it has become extremely difficult to manage and track the user's behavior in social media. Malicious actors leverage this weakness of social media platforms and target regular users with threats that include cyberbullying, malware distribution, spam distribution, fake news, and propaganda. The consequence of such threats can affect a large number of people and can result in catastrophic damage. However, identifying the malicious users from the huge number of regular users remain the most challenging problem for the social media platforms. Big data analytics can be one of the most powerful tools for the social media platforms to prevent such attacks on social media platforms. This paper discusses the threats of social media and the ways to use big data analytics to prevent such attacks on social media platforms.

## KEYWORDS

E534, HID 237, Big Data, Social Media, Threat Intelligence, Privacy

## 1 INTRODUCTION

More than two billion people use various social media platforms every day [18]. In every minute, approximately one million people log into Facebook, 50,000 photos are uploaded on Instagram, half million tweets are posted on Twitter, two million snaps are created, and one million profile matches happen on Tinder (Figure 1) [8]. Social media platforms have become a virtual society for everyone where people are interacting with a large number of an audience every day. Similar to the regular society, there are bad actors in the virtual society who are trying to harm people. Due to the extended outreach, social media platforms have become an ideal platform for the malicious actors to harm that includes cyberbullying [5, 7, 9, 16, 17] and the distribution of offensive, misleading, false or malicious information [6, 10, 13, 14]. Terrorist and government can also leverage social media to spread propaganda [3, 20].

[Figure 1 about here.]

Since the beginning of society, malicious actors are common phenomena and society has taken necessary steps to control them. Law enforcement organizations have been helping the society to control social menaces and protect the individuals from internal and external threats. Real society constituted by small groups, therefore, it is easier to control them. In contrast to the actual society, it is far more difficult to manage the virtual society due to its volume. It is nearly impossible to construct virtual law enforcement organizations in the virtual world and protect the individuals from malicious actors. Therefore, protecting the social media platforms

from threats remain one of the most challenging problems. The recent advancement of machine learning and big data shows promises and offer a new set of weapons to fight. Big data analytics provides an easier way to scale, manage, and visualize user's data which can be valuable for fighting malicious actors. The volume of data gives the researchers tremendous insight which can be a new way to help regular users.

There are a plethora of risks that Social Medias are facing every day. However, some threats can impact the user's safety and security. For that reason, generally, these platforms invest significant resources to prevent that. As already discussed earlier, regular analytics might not be fruitful to prevent such thing at scale, big data analytics gives more power to the providers and shows significant promises. This paper discusses four such safety and security threats and discusses how big data analytics have been used to reduce the impact.

## 2 SOCIAL MEDIA THREAT INTELLIGENCE

This section discusses the impact of big data on social media threat intelligence. Each threat were discussed first then the use of big data on detecting and preventing the threats was discussed:

### 2.1 Cyberbullying detection and identification

Cyberbullying can be defined as the use of computing devices to hurt or embarrass another person. Cyberbullying in social media constitutes posting negative or offensive comments in posts, post videos or photos to make fun of others, stalking, harassing, and trolling. Approximately, 43 percent teenagers in the U.S have been victims of cyberbullying in 2013 and most of them were bullied in social media [11]. Cyberbullying has a bad impact on people, which include deep emotional trauma, mental disorder, substance abuse, and suicidal tendency [19].

The rise of social media has caused significant growth in cyberbullying. The rise of photos and social media have aggravated the situation. However, text analysis or social media post analysis has become impressively smart to detect cyberbullying [7]. Natural language processing algorithms like LDA can easily detect social media posts with negative meanings and keyword matching can be helpful to detect and identify the offensive keywords. Recent results show promising advancement in detecting cyberbullying and in future it would be far easier to control. Another potential approach to detect and identify cyberbullying is analyzing photos or videos.

### 2.2 Information Abuse detection

Although social media abuse falls in the category of cyberbullying, abuse can be different than cyberbullying. Social media abuse can

be defined as misusing user's personal information that does not necessarily harm the user but can break the level of mutual trusts between the abuser and the abused. For example, stealing one's content from their social media profile does not harm the user but breaks the mutual trust in the relationship. In social media, people connect with others by putting a level of trust on others. Sometimes, bad actors can use the information to impersonate the person on social media. Later, the impersonated account can be used to embarrass or demean the victim. This might not be necessary causing mental harm to the victim but misusing the trust. Such bad actors are pretty common in social media. Since the bad actor's are using genuine information, it is very difficult to detect them.

One approach to detecting the abusers is to monitor the user's activity. Most of the cases, such actors exhibit some common behavior such as sending friend requests to random people, infrequent usage, random or anomalous behavior and such other behaviors. However, it is not possible for people to monitor the activity of the users to detect the abusers. However, common patterns can be helpful to detect such anomalies and data mining algorithms like k-means can be used to detect such anomalies. Big data analytics can also be helpful to blacklist the social media abusers and additional manual research can be useful to detect the abusers and ban them. For example, Xiao et al. [21] utilized Apache Hive to build a fake profile detection system and using Hadoop streaming they monitored the newly registered dataset. Blacklisted users then sent for manual reviewing. Likewise, other social media platforms adopted similar approach to detect the abusers.

[Figure 2 about here.]

### 2.3 Identifying the fake news/misinformation

Various terrorist and government organization utilizes social media to spread their propaganda. Often, these organizations use fake news or false information to spread their agenda or to deceive people [4]. Since fake news is often hard to identify people often deceived by them and share the news. Due to its large volume of users, fake news can spread significantly fast and can impact the society. According to a survey conducted by pew research center, approximately one-fourth of the U.S. adults have shared fake news [1]. Analysts and evidence suggested that Russian government set up numerous fake accounts to spread dubious information regarding U.S. election 2016 and eventually impacted the U.S. election [1, 2]. Due to the significant impact on an organization, identifying fake news detection and prevention garnered significant attention from the researchers.

By leveraging new machine learning tools, now it has become easier to track the online spread of misinformation and detecting social bots [13, 14]. Fact checking using knowledge graph can check facts in nearly run-time which can be useful to identifying fake news quickly and prevent misinformation [15]. More works use machine learning approaches to detect and prevent misinformation.

### 2.4 Preventing Terrorism

Various terrorist organizations like Al-Qaeda and ISIS use social media platforms to communicate with their members. They often use social media platforms to recruit new members [12]. Due to the

volume, now the impact of terrorism can be catastrophic and the increasing number of terrorist attacks on U.S and European countries proves that terrorists are becoming successful. Terrorism existed before and it exists now. But, modern technology has become a powerful tool for the terrorist organizations to amplify the impact.

Counterterrorism organizations need massive surveillance to prevent the terrorists from harming people. Without the help of Big Data, it will be impossible to support such massive surveillance. Moreover, a proper design of surveillance tools can be useful to maintain the privacy of regular citizens. Big data analytics has significant contribution to make this happen. Data visualization tools can help the counterterrorism organizations to monitor such surveillance.

## 3 CONCLUSION

This paper briefly discusses the cybersecurity and safety risks on various social media and explored some existing Big data approaches to tackle such problem. New tools have already been successful to prevent and control various threats on social media, but it needs more research and additional tools. In near future, the threats can increase significantly and big data analytics need to be prepared for that to prevent future threats.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski for helping us with the instruction and resources that were required to complete this paper. We would also like to thank the associate instructors for being available on the course website all the time and helping us with their answers.

## REFERENCES

- [1] Benedict Carey. NYTimes. 2017. How Fiction Becomes Fact on Social Media. <https://journalistsresource.org/studies/society/internet/fake-news-conspiracy-theories-journalism-research>. (2017). Online; accessed Oct 29, 2017.
- [2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. <https://doi.org/10.1257/jep.31.2.211>
- [3] Jessikka Aro. 2016. The cyberspace war: propaganda and trolling as warfare tools. *European View* 15, 1 (01 Jun 2016), 121–132. <https://doi.org/10.1007/s12290-016-0395-5>
- [4] Christoph Aymanns, Jakob Foerster, and Co-Pierre Georg. 2017. Fake News in Social Networks. *CoRR* abs/1708.06233 (2017). arXiv:1708.06233 <http://arxiv.org/abs/1708.06233>
- [5] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1217–1230. <https://doi.org/10.1145/2998181.2998213>
- [6] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Comm. ACM* 59, 7 (2016), 96–104. <https://doi.org/10.1145/2818717> Preprint arXiv:1407.5225.
- [7] Homa HosseiniMardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakanan Mishra. 2015. Detection of Cyberbullying Incidents on the Instagram Social Network. *CoRR* abs/1503.03909 (2015). arXiv:1503.03909 <http://arxiv.org/abs/1503.03909>
- [8] Jeff Desjardins. Visual Capitalist. 2017. How Fiction Becomes Fact on Social Media . <http://www.visualcapitalist.com/happens-internet-minute-2017/>. (2017). Online; accessed Oct 27, 2017.
- [9] Grace Chi En Kwan and Marko M. Skoric. 2013. Facebook bullying: An extension of battles in school. *Computers in Human Behavior* 29, 1 (2013), 16 – 25. <https://doi.org/10.1016/j.chb.2012.07.014> Including Special Section Youth, Internet, and Wellbeing.
- [10] Filippo Menczer. 2016. The Spread of Misinformation in Social Media. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. International World Wide Web Conferences

- Steering Committee, Republic and Canton of Geneva, Switzerland, 717–717. <https://doi.org/10.1145/2872518.2890092>
- [11] National Crime Prevention Council. 2014. Stop bullying before it starts. <http://www.ncpc.org/resources/files/pdf/bullying/cyberbullying.pdf>. (2014). Online; accessed Oct 27, 2017.
- [12] Operation 250. 2017. How Terrorists use the Internet? <https://www.operation250.org/how-terrorists-use-the-internet/>. (2017). Online; accessed Oct 27, 2017.
- [13] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proc. 25th International Conference Companion on World Wide Web*. <https://doi.org/10.1145/2872518.2890098> Preprint arXiv:1603.01511.
- [14] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. *The spread of fake news by social bots*. Preprint 1707.07592. arXiv. <https://arxiv.org/abs/1707.07592>
- [15] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2017. Finding Streams in Knowledge Graphs to Support Fact Checking. In *Proc. IEEE International Conference on Data Mining (ICDM)*. <https://arxiv.org/abs/1708.07239>
- [16] Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. 2017. "They Basically Like Destroyed the School One Day": On Newer App Features and Cyberbullying in Schools. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1210–1216. <https://doi.org/10.1145/2998181.2998279>
- [17] Robert Slonje, Peter K. Smith, and Ann Frisen. 2013. The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior* 29, 1 (2013), 26 – 32. <https://doi.org/10.1016/j.chb.2012.05.024> Including Special Section Youth, Internet, and Wellbeing.
- [18] Statista. 2017. Most famous social network sites worldwide as of September 2017, ranked by number of active users (in millions). <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. (2017). Online; accessed Oct 27, 2017.
- [19] VeryWell. 2017. What Are the Effects of Cyberbullying? Discover how cyberbullying can impact victims. <https://www.verywell.com/what-are-the-effects-of-cyberbullying-460558/>. (2017). Online; accessed Oct 27, 2017.
- [20] Gabriel Weimann. 2006. *Terror on the Internet: The New Arena, the New Challenges*. The United States Institute of Peace.
- [21] Cao Xiao, David Mandell Freeman, and Theodore Hwa. 2015. Detecting Clusters of Fake Accounts in Online Social Networks. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (AISeC '15)*. ACM, New York, NY, USA, 91–101. <https://doi.org/10.1145/2808769.2808779>

W

## 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

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

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

Other formatting issues

### A.4 Writing Errors

DONE:

Errors in title, e.g. capitalization

DONE:

Spelling errors

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

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 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

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

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 inflate 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 float. For this class, you should place all figures at the end of the report.

DONE:

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

DONE:

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

DONE:

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

DONE:

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

e.g.

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

#### LIST OF FIGURES

1	What happens in an internet minute? [8]	6
2	Fake User detection approach used by LinkedIn[21]	6

# 2017 This Is What Happens In An Internet Minute

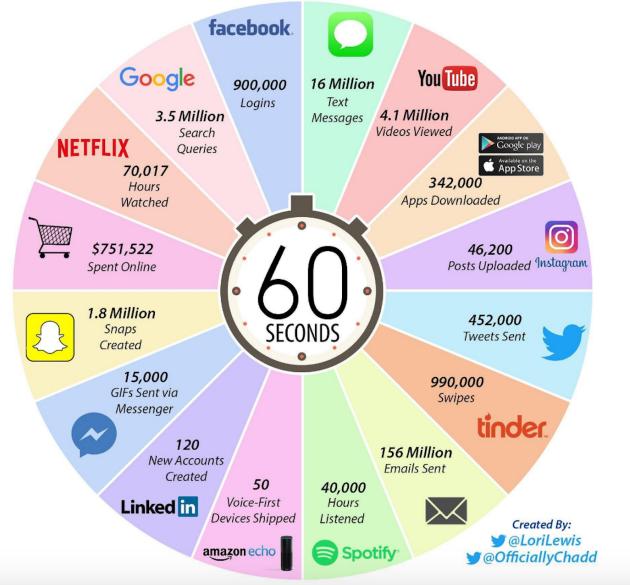


Figure 1: What happens in an internet minute? [8]

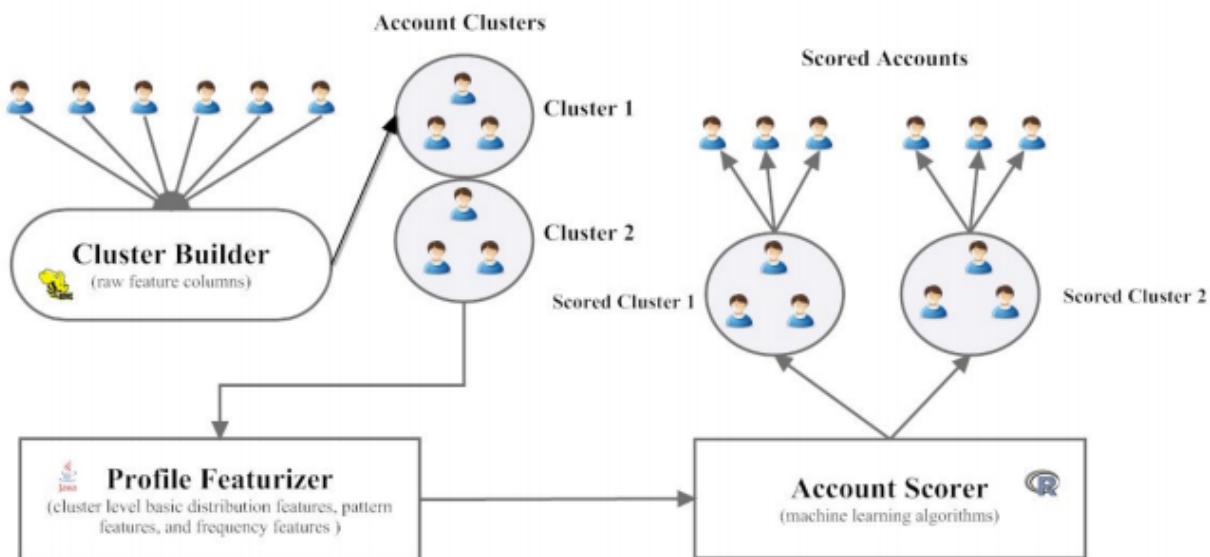


Figure 2: Fake User detection approach used by LinkedIn[21]

```
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--page numbers missing in both pages and numpages fields in Aymanns:2017
Warning--numpages field, but no articleno or eid field, in Cheng:2017
Warning--page numbers missing in both pages and numpages fields in HosseiniMardiMRH15
Warning--numpages field, but no articleno or eid field, in Menczer:2016
Warning--empty publisher in Shao15hoaxy
Warning--empty address in Shao15hoaxy
Warning--page numbers missing in both pages and numpages fields in Shao15hoaxy
Warning--empty publisher in Shiralkar2017Finding-Streams
Warning--empty address in Shiralkar2017Finding-Streams
Warning--page numbers missing in both pages and numpages fields in Shiralkar2017Finding-
Warning--numpages field, but no articleno or eid field, in Singh:2017
Warning--empty address in Weimann:2006
Warning--numpages field, but no articleno or eid field, in Xiao:2015:DCF:2808769.2808779
(There were 13 warnings)
```

```
bibtext _ label error
```

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

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

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

```
[2017-11-06 17.36.49] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
```

```
bookmark level for unknown defaults to 0.  
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.  
Typesetting of "report.tex" completed in 1.7s.
```

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

```
name: Ahmed, Tousif  
hid: 237  
paper1: 100%, October 27, 2017  
paper2: 99%, In Progress  
project: in progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6  
wc 237 paper2 6 1744 report.tex  
wc 237 paper2 6 2857 report.pdf  
wc 237 paper2 6 1215 report.bib
```

```
find "
```

---

```
passed: True
```

---

```
find footnote
```

---

```
passed: True
```

---

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

---

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

```
passed: True
```

---

```
floats
```

---

38: More than two billion people use various social media platforms every day~\cite{social-media}. In every minute, approximately one million people logs into Facebook, 50,000 photos are uploaded on Instagram, half million tweets are posted on Twitter, two million snaps are created, and one million profile matches happen on Tinder (Figure ~\ref{f:socialmedia}) ~\cite{social-media2}. Social media platforms have become a virtual society for everyone where people are interacting with a large number of an audience every day. Similar to the regular society, there are bad actors in the virtual society who are trying to harm people. Due to the extended outreach, social media platforms have become an ideal platform for the malicious actors to harm that includes cyberbullying ~\cite{Sonje:2013,Kwan:2013,Singh:2017,Cheng:2017,HosseiniMardiMRH15} and the distribution of offensive, misleading, false or malicious information ~\cite{Menczer:2016, socialbots-CACM, Shao15hoaxy, Shao17hoaxybots}. Terrorist and government can also leverage social media to spread propaganda~\cite{Aro2016, Weimann:2006}.

41: \begin{figure}[!ht]  
42: \centering\includegraphics[width=0.5\columnwidth]{images/one-  
internet-minute.png}  
43: \caption{What happens in an internet minute?~\cite{social-  
media2}}\label{f:socialmedia}  
63: \begin{figure}[!ht]  
64: \centering\includegraphics[width=\columnwidth]{images/fakeuser.png  
}  
66: \label{f:fake}

figures 2  
tables 0  
includegraphics 2  
labels 2  
refs 1  
floats 2

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

Label/ref check  
passed: True

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

```
find textwidth
```

---

```
passed: True
```

---

```
below_check
```

---

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

```
This is BibTeX, Version 0.99d (TeX Live 2016)
```

```
The top-level auxiliary file: report.aux
```

```
The style file: ACM-Reference-Format.bst
```

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

```
Warning--page numbers missing in both pages and numpages fields in Aymanns:2017
```

```
Warning--numpages field, but no articleno or eid field, in Cheng:2017
```

```
Warning--page numbers missing in both pages and numpages fields in HosseiniMardiMRH15
```

```
Warning--numpages field, but no articleno or eid field, in Menczer:2016
```

```
Warning--empty publisher in Shao15hoaxy
```

```
Warning--empty address in Shao15hoaxy
```

```
Warning--page numbers missing in both pages and numpages fields in Shao15hoaxy
```

```
Warning--empty publisher in Shiralkar2017Finding-Streams
```

```
Warning--empty address in Shiralkar2017Finding-Streams
```

```
Warning--page numbers missing in both pages and numpages fields in Shiralkar2017Finding-
```

```
Warning--numpages field, but no articleno or eid field, in Singh:2017
```

```
Warning--empty address in Weimann:2006
```

```
Warning--numpages field, but no articleno or eid field, in Xiao:2015:DCF:2808769.2808779
```

```
(There were 13 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
```

# Prediction of psychological traits based on Big Data classification of associated social media footprints

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

## ABSTRACT

Discusses the importance of digital footprints in evaluating person's psychological traits. We also reviewed few researches and articles which conducted studies in this field. We presented an algorithm at very high level of abstraction to understand how digital qualitative data can be translated to quantitative data to arrive at psychological traits. We concluded by providing few real life examples such as how Facebook likes can be used to evaluate psychological traits, how this research was used in last year elections and etc.

## KEYWORDS

Big Data, Edge Computing i523, psychological traits, Big Data, Facebook Data, Social media, digital foot prints, five factor model, personality traits, elections, Facebook likes, Facebook comments, Instagram

## 1 INTRODUCTION

With the advancement of digital media and social media networks, there has been enormous amount of human activities, which is recorded as the digital footprints. According to IBM, in 2012 on an average 500 MB of personal data is uploaded to the online digital database daily. This data is either in the form of social media activities such as Facebook likes, Facebook comments, profile picture upload, tweets or in the form offline transactions where person goes to grocery shopping and pays using credit card. According to [6] China is investing heavy technological resources to mine this data along with person's financial transactions to build social credit system. This project is expected to be implemented by 2020. There has been studies [1] – [12], which analyzed the behavior outcomes of the digital profile with the actual characteristics of an individual. Interesting thing about these studies is that human behavior can be mapped statistically to define similarities and differences between individuals. This can further be used to build recommendation based system to enrich social medial networks such as Facebook, LinkedIn, and Twitter etc. These studies [1] to [12] further contributes in radically improving our behavior understanding of humans. [8] discusses about the predictability of individual's psychological traits using statistical approach to arrive at the personality traits with certain confidence level. Psychological traits automation can further be used to enrich the quality of recommendation based systems and online search engines. [3] suggest how these studies [1] and [12] can be used to improve online marketing systems. With so many advantages on one side, on other side it possesses biggest challenge to the Data privacy [2] and [10]. Reason why these studies [1] and [12] provide better estimate of

human psychological traits as compared to results of psychometric test because these study results [1] and [12] takes the data of prolonged history. However, psychometric tests on the other hands is for few minutes or hours where human can manipulate response in order to achieve desire results. Thus, these studies [1] and [12] can also be leveraged in employee hiring process where many companies still relies on psychometric tests.

## 2 DATA SOURCE OF BIG DATA IN DIGITAL WORLD

This section discusses how we can import, store and preprocess digital big data. This data can be fetched online via REST api or its direct available to download from website such as mypersonality.org. This site stores the social media data of close to six million participants. There are other sites like Stanford network analysis project, which contains enormous amount of data in the form of product reviews, Tweets, and social media data. Social medial sites like Instagram and Twitter provides public rest APIs through which we can access data, which is public. Other example is Amazon.com, which provides elegant web services to access product reviews. For preprocessing of this data, Python provides excellent libraries to access [via web service call] and preprocess data.

## 3 HUMAN BEHAVIOR AND PERSONALITY

[11] talks about various models, which can be used to describe human personality. Among all, five factor model [FFM] is proved to be the best model to describe human behavior, psychological traits and preferences: Openness, Conscientiousness, Extroversion, Agreeableness and Emotional stability. We have data, we have psychological traits, and biggest challenge lies in extracting value out of big data and mapping the result to psychological traits. To accomplish this challenge we can perform singular value decomposition to map the qualitative data to quantitative data. To elaborate this further let us take an example: we have a Facebook likes of 10 million people and we filter down top 100 Facebook pages, which are of relevance. Top 100 relevant pages are those, which can predict factors mentioned in FFM. Now we will prepare Boolean matrix with Facebook user profile on vertical axis and Facebook page as horizontal axis. In simple words row represents Facebook user and column represents Facebook page. We will mark the coordinate as one if corresponding Facebook user [on vertical axis] likes a page [on horizontal axis] otherwise zero. Therefore, matrix will look like this:

	<i>fbPage<sub>1</sub></i>	<i>fbPage<sub>2</sub></i>	...	<i>fbPage<sub>n</sub></i>
<i>user<sub>1</sub></i>	1	0	...	1
<i>user<sub>2</sub></i>	0	1	...	1
<i>user<sub>3</sub></i>	:	:	..	:
<i>user<sub>n</sub></i>	1	1	...	1

These 100 Facebook pages is clustered, based on the five factors mentioned in FFM. First twenty pages will represent first factor, second twenty pages will represent second factor and so on. Next step would be to build correlation matrix that represents how each person is correlated with each other based on the five factors. This matrix will be N by N where is N is number of Facebook users in this experiment. This matrix will help to determine how similar Facebook users are. Which will help us to build the recommendation based systems because similar peoples tends to like same pages and share same psychological traits. This correlation matrix will look like this:

	<i>user<sub>1</sub></i>	<i>user<sub>2</sub></i>	...	<i>user<sub>n</sub></i>
<i>user<sub>1</sub></i>	1	.75	...	.85
<i>user<sub>2</sub></i>	.75	1	...	.91
<i>user<sub>3</sub></i>	:	:	..	:
<i>user<sub>n</sub></i>	.85	.91	...	1

- 
- Step 1:** Build binary matrix with Facebook user profile on vertical axis and Facebook page as horizontal axis.
- Step 2:** Populate the binary matrix with one and zero depending on if person has liked the page or not.
- Step 3:** Sort Facebook page columns depending on the factors mentioned in FFM.
- Step 4:** Use this matrix to build correlational matrix represents how each person is correlated with each other based on the five factors.
- Step 5:** Apply k mean algorithm to group Facebook users of similar factors mentioned in FFM.
- 

#### 4 COMPUTER BASED PERSONALITY JUDGMENT AND HUMAN BASED PERSONALITY JUDGMENT

Research [14] has shown computer based personality judgments are more accurate than those made by humans. According to [14] perceiving and judging people's personality is an important component of living society. Many cognitive decision made by humans are based on the judgment they have in their mind. This research [14] has shown how advance machine learning algorithms and statistical tools can be used to predict the personality traits and compared the results with the human judgments. This research also addresses the issue of substantiating the qualitative aspects of behavior with the quantitative parameters. Computer based personality judgment is not only based on machine learning or statistics but computer vision algorithms can also be used to distinguish facial emotions and concluding psychological traits.

#### 5 SOCIAL NETWORK AS A PERSONALITY TRAIT PREDICTOR

[9] studies suggest how valuable social network is in predicting the psychological traits. According to [9], It is considered as one of the valuable digital footprints to predict intimate personal traits. For instance, number of friends and their location can be used to grade first factor of FFM, which is openness. Person romantic partner can be detected depending on the social network overlap of each friend, which can further be analyzed to predict one's sexual preference. These predictions can further be statistically analyzed to [14] to know how accurate predictions are. We can use social network data on the algorithm discussed in the "Human Behavior and Personality" and conclude a very strong predictions on the psychological traits of a person. It has been in the news that 2016 elections were strategized with the help of the social media big data which will be discussed in the next section.

#### 6 SOCIAL MEDIA BIG DATA AND ITS IMPACT ON POLITICAL ELECTIONS

[13] suggests how last year elections were revolutionized by the impact of big data of social media. Using statistical and machine learning algorithms on social media big data, political parties filtered down the data to identify their likely supporters and then channelized their strategy to win their votes. These strategies were less expensive than conducting campaigns at various places. Traditional analysis is generally based on the survey which is in the sense is limited [7] but now with the ease of big social media data, analysis is more accurate and conclusive. There has been sophisticated tools available that can predict the person's race depending on his or her name and location. In recent election, political parties also combined social media data and public data [from census Bureau] to run sophisticated machine learning algorithm to pinpoint their supports. All these mentioned ways helped the political parties to micro target their supporters and gained their votes.

#### 7 SOCIAL ACTIVITY, THE PREDICTOR OF PERSONALITY

[9] suggests Facebook profile of a user is not static rather it also contains enriched records of digital footprints such as likes, comments, reactions to other posts. Such activities materializes the connections between user and content. This information along with the other activities such as playlist, browsing logs, online shopping activities and google queries can be used to develop sophisticated highly predictive FFM set for a user and with a very high confidence level can predict user's age, gender, intelligence religious view and sexual orientation [9]. Very interesting example from the [9] suggests "Users who liked Hello Kitty brand tended to have high openness, low conscientiousness, and low agreeableness" - strange but very interesting! [9] research further elaborate the importance of comments. Semantic analysis on comment can be analyzed to infer one's personality as shown by the research: [5] and [4].

#### 8 CONCLUSION

We discussed various ways in which social medial data can be utilized to build five factor personality model for a user. Main

purpose here is to review the literature work done in this field and also presented the algorithm which can be used to translate qualitative data to quantitative data and how value can be extracted to build FFM for a user. We discussed computer based personality judgments are better than the human based personality judgments. We also touched based where social network can be used to predict user's personality. As discussed earlier, these researches [1] - [12] have proved to impact the general election last year in United States. Finally we concluded by showing evidences how social activity can be used to build the FFM for a user.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and all the TA's for their support and suggestions to write this paper.

## REFERENCES

- [1] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 445–454. <https://doi.org/10.1145/2030112.2030171>
- [2] Declan Butler. 2007. Data sharing threatens privacy. *NCBI* 449 (11 2007), 644–5.
- [3] Ye Chen, Dmitry Pavlov, and John F. Canny. 2009. Large-scale Behavioral Targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, New York, NY, USA, 209–218. <https://doi.org/10.1145/1557019.1557048>
- [4] Adam D. I. Kramer and Kerry Rodden. 2008. Word usage and posting behaviors: Modeling blogs with unobtrusive data collection methods. (01 2008), 1125–1128 pages.
- [5] Samuel Gosling, Sam Gaddis, and Simine Vazire. 2007. Personality Impressions Based on Facebook Profiles. *ICWSM* 7 (Jan. 2007), 1–4.
- [6] Lucy Hornby. 2017. China changes tack on fiscal credit scheme plan. eNewsPaper. (July 2017). <https://www.ft.com/content/f772a9ce-60c4-11e7-91a7-502f7ee26895> China changes tack on fiscal credit scheme plan.
- [7] Sean Illing. 2017. A political scientist explains how big data is transforming politics. vox. (March 2017). <https://www.vox.com/conversations/2017/3/16/14935336/big-data-politics-donald-trump-2016-elections-polarization>
- [8] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. <https://doi.org/10.1073/pnas.1218772110> arXiv:<http://www.pnas.org/content/110/15/5802.full.pdf>
- [9] Renaud Lambiotte and Michal Kosinski. 2014. Tracking the Digital Footprints of Personality. *IEEE* 102 (12 2014), 1934–1939.
- [10] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. (06 2008), 111–125 pages.
- [11] Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist* 48 (02 1993), 26–34.
- [12] Kern ML Dziurzynski L Ramones SM Agrawal M Shah A Kosinski M Stillwell D Seligman ME Ungar LH Schwartz H, Eichstaedt JC. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. (2013). <https://www.ncbi.nlm.nih.gov/pubmed/24086296>
- [13] Chuck Todd and Carrie Dann. 2017. How Big Data Broke American Politics. eNewsPaper. (March 2017). <https://www.nbcnews.com/politics/elections/how-big-data-broke-american-politics-n732901> How Big Data Broke American Politics.
- [14] Youyou Wu, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *PNAS* 112 (01 2015), 1–5.

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 this change TODO with DONE. However if you check something on with DONE, but we find you actually have not executed it correctly,

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.

## 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

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 float. 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
```

## 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 inflate 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

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 "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst





# latex report

```
38:89    error    trailing spaces  (trailing-spaces)
39:81    error    line too long (131 > 80 characters)  (line-length)
39:131   error    trailing spaces  (trailing-spaces)
40:81    error    line too long (130 > 80 characters)  (line-length)
40:130   error    trailing spaces  (trailing-spaces)
41:62    error    trailing spaces  (trailing-spaces)
42:25    error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Arora, Gagan
hid: 301
paper1: 100% Oct 29 17
paper2: 100% Nov 4
project: in progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
4
wc 301 paper2 4 2007 report.tex
wc 301 paper2 4 2928 report.pdf
wc 301 paper2 4 1492 report.bib
```

```
find "
```

---

```
95: \cite{ref13} studies suggest how valuable social network is in
predicting the psychological traits. According to \cite{ref13},
It is considered as one of the valuable digital footprints to
predict intimate personal traits. For instance, number of friends
and their location can be used to grade first factor of FFM, which
is openness. Person romantic partner can be detected depending on
the social network overlap of each friend, which can further be
analyzed to predict one\textquotesingle s sexual preference. These
predictions can further be statistically analyzed to \cite{ref12}
to know how accurate predictions are. We can use social network
data on the algorithm discussed in the "Human Behavior and
```

Personality" and conclude a very strong predictions on the psychological traits of a person. It has been in the news that 2016 elections were strategized with the help of the social media big data which will be discussed in the next section.

103: \cite{ref13} suggests Facebook profile of a user is not static rather it also contains enriched records of digital footprints such as likes, comments, reactions to other posts. Such activities materializes the connections between user and content. This information along with the other activities such as playlist, browsing logs, online shopping activities and google queries can be used to develop sophisticated highly predictive FFM set for a user and with a very high confidence level can predict users age, gender, intelligence religious view and sexual orientation \cite{ref13}. Very interesting example from the \cite{ref13} suggests "Users who liked Hello Kitty brand tended to have high openness, low conscientiousness, and low agreeableness" - strange but very interesting! \cite{ref13} research further elaborate the importance of comments. Semantic analysis on comment can be analyzed to infer one's personality as shown by the research: \cite{ref16} and \cite{ref17}.

passed: False

find footnote

---

passed: True

find input{format/i523}

---

5: \input{format/i523}

passed: True

floats

---

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

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

Label/ref check  
passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

Database file #1: report.bib

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst

Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3085 of file ACM-Reference-Format.bst





Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "Schwartz H, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones while executing---line 3229 of file ACM-Reference-Format.bst  
(There were 64 error messages)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

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

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

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

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

```
find newline
```

```
-----  
83: \textbf{\textit{Step 1}}: Build binary matrix with Facebook user  
profile on vertical axis and Facebook page as horizontal  
axis.\newline
```

```
84: \textbf{\textit{Step 2}}: Populate the binary matrix with one and  
zero depending on if person has liked the page or not.\newline
```

```
85: \textbf{\textit{Step 3}}: Sort Facebook page columns depending on  
the factors mentioned in FFM.\newline
```

```
86: \textbf{\textit{Step 4}}: Use this matrix to build correlational  
matrix represents how each person is correlated with each other  
based on the five factors.\newline
```

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

```
find cite {
```

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

# Big Data Applications in Historical Studies

Neil Eliason

Indiana University

Anderson, Indiana

## ABSTRACT

As big data analytics progress in other fields, historians have began to consider how they can apply these techniques to their studies. Various studies demonstrate potential benefits of big data approaches. However, care must be taken to keep big data results in the overall context of traditional scholarship and to utilize appropriate historical and technical expertise to avoid introducing inaccuracy and bias into findings.

## KEYWORDS

i523, HID312, Big Data, History, Data Visualization, Inter-disciplinary

## 1 INTRODUCTION

### 1.1 Big Data

To date big data can claim numerous victories in a variety of fields, and promises more. Businesses such as Facebook and Netflix have built corporate empires off of the insights gathered from their big data, and physicists and biologists are learning what makes up the universe and ourselves via big data [1].

Despite all this, the concept itself is rather nebulously defined. A rough description is data with quantitative factors that require specialized techniques to utilize. The most commonly referenced big data factors are volume (amount of data), variety (number of data source types), and velocity (rate of data collection or input) known as “the three vs.” As these data factors become more extreme, to the point that traditional methods of data analysis fail, it becomes big data. While this definition is generally accepted, its application varies based upon the industry or field of study and often changes with developments in information technology [5].

The focus on big data arises partially from the phenomenon of data storage capabilities growing at a faster rate than data processing. This creates a situation where data can be economically stored, but not as economically processed, requiring specialized analytic techniques. As big data progresses through the storage, cleaning, analysis, and interpretation stages of the data life cycle, specialized approaches are required [1].

### 1.2 History of History

The historian’s labor has involved interacting with voluminous and varied data for centuries. Before computers, this process involved searching physical archives for relevant data, and manually copying and organizing it into useful information to be analyzed. Though this method can deliver deep insights, some data sets are too big to be studied in a manual fashion [7].

Around the mid-twentieth century, computers became sufficiently powerful and usable for historians to begin using them to process larger amounts of information. This facilitated a change towards a more quantitative approach to historical analysis and

a focus by some from tracing the rise and fall of political or ideological forces, to developing a more complete understanding of mundane topics, such as the family or economics.

As archives become digitized and accessible via the internet, the quantity of data available leads to an appeal to big data analytic methods [4]. The potential of unlocking significant connections and developing big picture historical insights at the scale of the growing digital archives of the world is alluring. This hope has driven the labor of many researchers towards developing more big data informed research methods and has directed funds of many institutions towards investments in data infrastructure. However, many are also concerned that the promises of big data are at best optimistic, and at worst hiding potential pitfalls to the historical process [7].

### 1.3 Thesis

Big Data Analytics have the potential to provide new insights to the field of historical studies. However, their application will differ due to the nature of historical data, and they will serve as an additional tool for the historian, rather than replacing more traditional approaches.

## 2 BIG DATA IN HISTORICAL STUDIES

### 2.1 Data Sources

It could be argued that history has had big data for some time, but that the lack of computational capability prevented it from being accessed on a large scale. As big data analytics mature, pressure develops to increase the data available for analysis by digitizing more archival material. This is evidenced not only by the familiar repositories of e-books, but also by archives of a variety of types, such as newspapers articles [7] or letters [4].

Sources for big data research consist not only of the content of documents in an archive, but also the bibliographical records. While originally designed to allow individual works to be located in an archive, historians have began to study the bibliographical data themselves, an approach called distant reading. By looking at the data about a document, rather than the document’s content, societal or intellectual trends can be identified across large scale factors such as time or geography in a more comprehensive way. This approach has elicited some criticism that collections of bibliographical data are not complete enough to derive such large-scale conclusions. Still, considerable interest exists in targeting these data sets for historical analysis [10].

However, the data from these sources differs from that of other fields which utilize big data analytics. Historical data is not streaming the way that social media or smartphone sensors are. It is data which has already been collected, organized, and often times analyzed for a purpose defined by people from a different time and different needs/constraints from ourselves. This creates data

sets which are difficult to compare and often require considerable cleaning and reworking to be used in a larger framework. [4].

## 2.2 Analytics for Big Historical Data

Due to the natural reliance on documents in historical studies, text analytic techniques are the primary set of big data approach utilized by historians. Text analytics are a broad category of related algorithms and statistical techniques, such as artificial intelligence, machine learning, and natural language processing that attempt to extract specific information from the text and identify patterns and relationships within the body of data [7].

Artificial intelligence is “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings” [2]. In the context of historical research, this would include tasks such as extracting relevant content from sources or identifying relationships within the data. A specific type of artificial intelligence is machine learning, which consists of programs which change their actions autonomously in response to external input. Their ability to adapt allows them to do decision-making tasks, and thus can search through data sources in a more intelligent way to find relevant data [1]. Natural language processing is another artificial intelligence technique, which aims to create programs that can take human language, and make it machine readable [9]. Historians can use such programs to extract meaningful information from archival documents and prepare it for more further analysis and interpretation.

In order to interpret the results of big data analysis, visualization is critical. This is a challenge, as the large scale of the data makes striking a balance between a sufficiently big picture perspective without losing relevant details difficult. Many approaches attempt to utilize high resolution approaches to avoid losing important information [1]. This process is especially challenging in historical studies, as the data is often incomplete and may have inconsistencies which prevent assuming a uniform set of data. For this reason, historians often use visualizations to identify qualitative, rather than quantitative relationships in the data, to inform further inquiry [4].

## 2.3 Software Packages and Resources for Big Data History

A variety of software packages have been utilized to assist the process of translating raw data into historical insights, such as Tableau, Gephi, R, and ArcGIS. However, a limitation of these tools is their quantitative focus, which tends to exclude more qualitative approaches [4]. Some general qualitative analysis software has been applied to big data historical analysis, such as Google Fusion Tables and OpenHeatMap [10].

Some software has been developed to provide a more qualitative visualization tool set for researchers. For example Stanford University developed a software package called Palladio, designed to visualize connections in large scale historical data. Their approach focused on visualizations that encouraged exploring data, rather than creating statistical statements about it. Examples of this would be mapping connections between historical actors over geography or creating a visualization of the social network of a particular figure in history. They do not create statistical arguments, rather they

give a framework for understanding how the data are connected [6].

Another tool with a qualitative visualization focus is the Web Application for Historical Sentiment analysis on Public media or WAHSP. Its specific purpose is to conduct text analysis on the National Library of the Netherlands digitized newspaper collection, which contains around 100 million articles published in 1618 to 1995. It provides a number of useful analyses, such as word frequency cloud visualizations, detecting positive or negative sentiment related to certain terms, and Named Entity Recognition, which can identify people, places, events, etc. and then connect them into a relational or geographical framework. It also provides an interactive histogram where the resolution of the data can be adjusted to quickly move between a big picture and detailed data perspective. A derivative project is BILAND, which is a program developed by Utrecht University, that builds off of WAHSP’s analytical capabilities, but adapts them across the Dutch and German languages for comparative cultural studies [7].

Along with these data intensive tools specifically designed for historical studies, there are also resources to help the historian learn some of these methods. For example, The Programming Historian website provides a wide range of tutorials and lessons on how to use digital tools in historical studies. At the time of this writing there were 67 lessons available organized by their target stage of research, including lessons on using R, Python, Java, and GitHub for historical studies[8].

## 2.4 Insights from Big Historical Data

A number of studies have used these techniques to approach historical research from a big data perspective. Stanford’s Mapping of the Republic of Letters project sought to map the social network of Enlightenment thinkers who actively corresponded with each other. This was accomplished by utilizing big data analytics on the meta-data of these letters to see how these thinkers related temporally, geographically, and socially. Through the research process, the need for more qualitative approaches to visualization was recognized, and eventually led to the development of the Palladio tool set.

Their analysis revealed a number of interesting points. By mapping the social network of John Locke, they supported previous scholarly contentions that the Enlightenment culture was not homogeneously connected, but was made up of a number of subcultures which had thin social connections. Also, by analyzing Benjamin Franklin’s letters, they noted that despite his reputation as cross cultural traveler, the main hub of his correspondence was between the familiar British cultural hubs in Philadelphia and London [4].

Another study used the WAHSP tool to research attitudes found towards drugs in early 20th century newspapers. It found by using the word cloud analysis tool, that before 1924 drugs such as heroin and opium were discussed in the context of health, but after 1924 they were more associated with crime. Their analyses also noted that Dutch negative associations with opium influenced their perception of China and the Dutch East Indies Colonies.

The related tool BILAND was used by to study how the perceptions of eugenics differed in the Netherlands and Germany, requiring an application which could compare data across languages.

The aim was to study not only the direct conversations about this topic in both regions, but also to study implicit use of terminology which was influenced by the eugenics debate. Through word cloud analysis, the study found that in the mid 19th century, eugenics and concepts of genetic inheritance were used in a primarily medical or biological context. By the 1930s, the terms were utilized more in reference to race and law [7].

One study analyzed music bibliographical data from the British Library and the Répertoire International des Sources Musicales to explore how music was transmitted in Europe over time and geography. Their analytic methods were actually closer to traditional techniques, using a large amount of research assistants to perform repetitive tasks, and wrestling with the information in Excel spreadsheets, but used visualization approaches more congruent with big data. They had surprising results related to who were the prominently published composers during different time periods. For example, during the 1800s, relatively unknown composers are high in the frequency list, and famous composers such as Bach did not make the top 50 [10].

### 3 POTENTIAL ISSUES

While big data can provide some powerful and at times novel solutions to problems, there are also potential issues with its implementation. For example as digital algorithms make search and selection decisions, bias can be introduced into the research inadvertently by the program. This danger is aggravated by the level of transparency of the algorithm, and how well the researcher understands it. For example, when researchers utilize commercial search engines, such as Google scholar, the algorithms are not available, and thus the researcher does not know why data is being included or excluded. If recommender systems are utilized, the potential for bias increases, as the search engine is actively attempting to provide results which are based on its user profile. This could exclude opportunities for data which may challenge the researcher's perspective. The danger of biased analysis through ignorant execution of an automated search or analysis is present in any big data tool, such as those previously described [3].

In the context of historical studies, it is acknowledged that to use digital methods without expert knowledge of both the subject matter and the big data methodologies can lead to inaccurate conclusions [4]. However, this can be addressed using a number of strategies. For example, there are resources to help historians expand their technical abilities, giving them greater understanding and control over big data analytic methods [8]. Creating inter-disciplinary teams are also an effective way to address biased analysis. By allowing information technology and historical research experts to meet together to create research methods, they can avoid unintentional bias from misuse of algorithms and from a lack of knowledge of historical context. However, equally important is for the research team to keep a balanced perspective on the role of big data analytics applied in historical studies. These new methods cannot be done in a vacuum or be used to replace traditional human reading of the sources [4]. Though big data techniques have powerful possibilities, they cannot replace the role of the historian, who combines their historical knowledge and narrative creation, to provide context and meaning to the enormous bits of information from the past [7].

There are also a number of technical difficulties associated with using big data for historical analysis. The big data available to historical researchers has no guarantee of completeness or uniformity from which to make generalized claims. Large archives of records can only provide information about what people in the past chose to record or which records survived to our time. Thus, traditional methods are critical, and big data methods serve the purpose of confirming or challenging previous theories, or inspiring new veins of inquiry. History is an interpretative task, and big data analytics serve to better inform interpretation, not replace it. In addition, data often comes from different sources formatted for a variety of purposes. Thus for the historian, rather than dealing with large masses of unstructured data, the challenge is to reconfigure data which has already been organized, and often is at cross-purposes to a researchers objectives [4].

### 4 CONCLUSION

Big data analytics have attracted both interest and criticism from historians. Large digitized databases, effective text analytic techniques, and innovative qualitative visualizations provide fertile ground for a big data approach to historical analysis, which would allow for a more comprehensive analysis of large data sets, which would not be possible for the researcher. These techniques have already been applied to a variety of topics, yielding useful, if not incredibly surprising results.

As historians continue to explore new methods of big data research, it is important they do so from a position of historical and technical expertise, to prevent inaccurate and biased findings. The researchers' perspectives on big data analysis also needs to remain balanced, not ignoring the possibilities of the new techniques, but also not neglecting traditional research. Without traditional scholarship, big data has no external validation or historical context, thus making its results inaccurate or meaningless.

### ACKNOWLEDGMENTS

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

### REFERENCES

- [1] C.L. Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, Supplement C (2014), 314 – 347. <https://doi.org/10.1016/j.ins.2014.01.015>
- [2] B.J. Copeland. 2017. artificial intelligence (AI). Webpage. (01 2017). <https://www.britannica.com/technology/artificial-intelligence>
- [3] Malte C. Ebach, Michaelis S. Michael, Wendy S. Shaw, James Goff, Daniel J. Murphy, and Slade Matthews. 2016. Big data and the historical sciences: A critique. *Geoforum* 71, Supplement C (2016), 1 – 4. <https://doi.org/10.1016/j.geoforum.2016.02.020>
- [4] Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. *Historical Research in a Digital Age*. *The American Historical Review* 122, 2 (2017), 400. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edssoaf&AN=edssoaf.a29ec0ac934f1257030b477fa5986b1cff6def96&ssite=eds-live&scope=site>
- [5] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 2 (2015), 137 – 144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- [6] Stanford Humanities and Design. 2017. Palladio. Visualize complex historical data with ease. webpage. (2017). <http://hdlab.stanford.edu/palladio/about/>

- [7] Eijnatten Joris van, Pieters Toine, and Verheul Jaap. 2013. Big Data for Global History: The Transformative Promise of Digital Humanities. *BMGN: Low Countries Historical Review*, Vol 128, Iss 4, Pp 55-77 (2013) 128, 4 (2013), 55. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsdjoj&AN=edsdjoj.6259f58bab47404485225cd4776fcf48&site=eds-live&scope=site>
- [8] Editorial Board of the Programming Historian. 2017. About the Programming Historian. Website. (10 2017). <https://programminghistorian.org/about>
- [9] Technopedia. 2017. Natural Language Processing (NLP). Webpage. (2017). <https://www.techopedia.com/definition/653/natural-language-processing-nlp>
- [10] Sandra1 Tuppen, Stephen2 Rose, and Loukia Drosopoulou. 2016. LIBRARY CATALOGUE RECORDS AS A RESEARCH RESOURCE: INTRODUCING 'A BIG DATA HISTORY OF MUSIC'. *Fontes Artis Musicae* 63, 2 (2016), 67 – 88. <http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=114128249&site=eds-live&scope=site>

## 5 ISSUES

DONE:

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

### 5.1 Assignment Submission Issues

DONE:

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

### 5.2 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

### 5.3 Formatting

DONE:

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

DONE:

Other formatting issues

### 5.4 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}

### 5.5 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

### 5.6 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

## 5.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

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

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
```

## 5.8 Details about the Figures and Tables

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

```
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
```

```
=====
[2017-11-06 17.37.23] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Label(s) may have changed. Rerun to get cross-references right.
Typesetting of "report.tex" completed in 1.6s.
./README.yml
 8:81      error    line too long (85 > 80 characters)  (line-length)
 9:81      error    line too long (87 > 80 characters)  (line-length)
 9:87      error    trailing spaces  (trailing-spaces)
10:80      error    trailing spaces  (trailing-spaces)
11:81      error    line too long (89 > 80 characters)  (line-length)
11:89      error    trailing spaces  (trailing-spaces)
12:81      error    line too long (85 > 80 characters)  (line-length)
25:81      error    line too long (94 > 80 characters)  (line-length)
25:94      error    trailing spaces  (trailing-spaces)
26:81      error    line too long (96 > 80 characters)  (line-length)
26:96      error    trailing spaces  (trailing-spaces)
```

```
27:81    error    line too long (97 > 80 characters)  (line-length)
27:97    error    trailing spaces  (trailing-spaces)
28:81    error    line too long (97 > 80 characters)  (line-length)
28:97    error    trailing spaces  (trailing-spaces)
29:81    error    line too long (83 > 80 characters)  (line-length)
37:55    error    trailing spaces  (trailing-spaces)
```

---

## Compliance Report

---

```
name: Neil Eliason
hid: 312
paper1: Review on 3 Nov 2017
paper2: 100%
project: not yet started
```

```
yamlcheck
```

---

```
wordcount
```

---

```
5
wc 312 paper2 5 2718 report.tex
wc 312 paper2 5 3548 report.pdf
wc 312 paper2 5 1064 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

passed: True

floats

---

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

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

Label/ref check

passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

The style file: ACM-Reference-Format.bst  
Database file #1: report.bib

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

=====

The following tests are optional

=====

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Analytics in Data Center Network Monitoring

Dhanya Mathew  
Indiana University  
711 N Park Ave  
Bloomington, Indiana 47408  
dhmathew@iu.edu

## ABSTRACT

Data Centers are evolving and adapting to newer strategies like virtualization. It is very challenging to monitor the current complex network infrastructure and its performance. Big data technologies promise solutions for network monitoring and performance analysis on real-time data. Big data streaming technologies offer high availability, high throughput, low latency and horizontally scalable solutions. Big data applications use distributed architectures and work on a huge volume of either offline data or streaming data or both.

Network monitoring solutions monitor the infrastructure, collect generated events, stream it and analyze it using a distributed analytics platform. Insights are derived out of this analysis. These insights facilitate the authority to take data-driven decisions. Big data analytics correlates data from different sources in real-time along with historical data to identify issues proactively. This helps either an automated system/human intervention to take necessary steps before the event actually happen. We explore the various network traffic monitoring and data analytics processes involved in the networking world like fault monitoring, performance monitoring and threat analysis.

## KEYWORDS

i523, HID328, big data, fault monitoring, threat analysis, event streaming, Flink

## 1 INTRODUCTION

Network administrators are mostly located remotely. Data center infrastructure monitoring and network traffic monitoring are one of the most critical parts of any enterprise. Good amount of planning should be there to choose the right monitoring solution, the set of things or events to be monitored and perform timely maintenance or upgrades of the tools [12]. Traditional Data Center monitoring tools can only monitor limited amount of events or thresholds. It is normally presented on a dashboard for the analyst to look on it but they could not relate the information from different sources. Often these tools gets outdated based on the high volume of data, newer technologies, data center scalability and configuration databases used [6].

A good monitoring solution needs to be able to alert you to issues with server hardware, operating system errors, application errors, networking hardware issues, and environmental issues. There is no single monitoring solution that can perform all of these functions. Hence an administrator should integrate multiple monitoring solutions to monitor different alerts and configure the alerts in such a way that it triggers email or message to the right team to take action. Als, the alerts need to be routed to a storage and analytical

solution as well for further analysis. Current monitoring solutions should be able to handle virtualization challenges [12].

By applying big data to data center operations, we can analyze the statistical and performance data obtained from multiple network devices like physical servers, virtual servers, routers, switches, firewalls, access points, storage etc. Big data can provide centralized and predictive analytics and we can identify the weak points in the system and determine what changes might improve the network performance [7].

[Figure 1 about here.]

As shown in Figure 1, data center monitoring platform is capable of monitoring all types of network devices. The events captured by the monitoring tools will be passed to the big data analytical platform for processing. Here events from different monitoring tools will be integrated and related to obtain the insights. These insights would be the base for deriving decisions and in turn results improvements in operational efficiency and financial performance etc.

## 2 DATA CENTER NETWORK MONITORING

There are a number of IT infrastructure monitoring tools which can be integrated with big data analytical platform. AppDynamics(Application and server performance monitoring), CopperEgg(IT Infrastructure monitoring), Datadog(Cloud monitoring), Logicmonitor(monitors networks, servers, storage and cloud), Gear5(Website performance monitoring) are some of them [9]. IT Infrastructure devices vary by type and manufacturer and the alarm or events need to be monitored will also vary accordingly. Fault, performance and SLA monitoring are very basic level of monitoring required for all types of devices.

### 2.1 Fault Monitoring

The fundamental task of system managers is to identify and rectify the faults in the network design and architecture. As today's huge Data Centers uses cloud networks, virtualization, parallel processing, load sharing etc, it is very crucial to detect, identify and remediate the network faults. Users may be connected to the network via locally or remotely via internet technologies. Faults in any of these areas will cause customer dissatisfaction.

Data centers are designed for scalability and hence network devices and servers are continuously get added, upgraded or replaced. Each fault that could happen in a data center can throw dozens of error reports. Hence the usage of a fault correlation software is important. This can be achieved by mechanisms like TL1 messages, SNMP traps, SYSLOG entries and application logs [8]. Below are the basic areas getting monitored as part of fault monitoring.

- Server and Network Alarms

- Server and Network Events
- Server and Network Event Enrichment
- Server and Network Automatic Notification

## 2.2 Performance Monitoring

Network performance depends on the type and capacity of the network connecting users to the application. Whenever an end user reports slow access to an application, the issue could be with the server or bandwidth or application itself [11]. An event correlation software is essential here as well.

Below are the basic areas getting monitored as part of performance monitoring.

- Network Performance
- Server and Network Performance Thresholds
- SLA Monitoring

## 3 DATA CENTER'S BIG DATA ANALYTICS

Bringing big data analytics to data center operations and can provide data-driven insights which may not be obtained from traditional monitoring tools. Infrastructure analytics tools are bringing big data processing techniques (e.g., Hadoop, NoSQL and Cassandra) into the data center, for quicker, more informed infrastructure management decisions [4].

Applications of big data are constantly growing and in turn the growth of data centers and cloud infrastructure as well. More than any of these, the number IoT devices have the most growth rate. According to Gartner Inc, by 2020, there will be 26 billion units IoT devices installed generating 44 Zetta Bytes of data in total [13], [5].

[Figure 2 about here.]

Figure 2 shows the data generation in billions per devices types. This huge data generation by default increases the growth of data centers by adding more and more servers and network devices. Even a simple addition to the current infrastructure would require detailed monitoring in terms of compliance, security and performance.

### 3.1 Server and Network Performance Data Analytics

Traditional data center monitoring tools are developed for pre-cloud era. Today, digital applications are distributed and traditional monitoring tools may not be effective. Hence Big Data based tools are necessary. Kentik introduced a new network performance management (NPM) tool for cloud based distributed applications and data centers [1]. Kentik's NPM solution builds on Kentik Detect, the big data-based, SaaS network analytics platform chosen by digital leaders like Yelp, Box, Neustar, Pandora and Dailymotion.

Kentik's NPM solution has a host based agent called nProbe which can be deployed in hybrid data centers and cloud networks. This could monitor and analyse network performance factors such as latency, retransmits, out of order packets, and packet fragments based on actual application traffic flows, offering the most relevant and actionable intelligence [1].

[Figure 3 about here.]

As shown in Figure 3, nProbe running on host computers send network performance monitoring data securely to Internet via encrypting proxy agent which is optional. Kentik data engine receives this data from internet and stores this data for actionable analytics. This data engine is big data based and can scale horizontally to store unsummarized data and provides powerful analytics for alerting, diagnostics and other use cases.

Big data performance analytics carried out on performance data are,

- Examine performance patterns
- Find commonalities
- Derive actionable insights

Below are the actionable insights that can be derived,

- Identifying devices exhibiting deviation from normal pattern
- Predict network performance
- Load balancing requirements

### 3.2 Server/Network Fault Data Analytics

In practice, historical failure data integrated with real-time data are often used to estimate the failure distribution of an item using statistical methods. This leads the way to predict future failures with a data-driven confidence level. Initially, this worked only when the particular device operated in a relatively stationary environment with no chances of changes. Given the complexity of modern systems, multiple failure mechanisms may interact with each other in a very sophisticated manner; environmental uncertainties may also have a great impact on the occurrence of failures. This requires predicting future failures of an item based on data which can reflect its real condition. It has been estimated that 99 percent of machinery failures are preceded by some malfunction signs or indications [14].

[Figure 4 about here.]

Figure 4, shows the network fault analysis architecture using big data. The procedure may include below activities as well.

- Mine network event patterns
- Find commonalities
- Actionable Insights and prediction of network Faults

### 3.3 Customer Care Data Analytics

Another main area of consideration in data center operations is Customer care. Using big data technologies we can track the entire customer journey including their previous and following operations. In particular related to data center, it is essential to analyze customer services, support emails, call logs and trouble tickets in order to understand their network and operational behaviour.

Based on bid data classification and clustering algorithms we can predict below actionable insights.

*Classification of Customers:*

- Customer Satisfaction Index
- Prediction of Customers to churn out of business

*Clustering of Customers:*

- New Customer Plans
- Network/Infrastructure Planning

### 3.4 Threat Analysis

Network is vulnerable to cyber-attacks. Once the network is compromised, the attacker can infiltrate the enterprise and gain access to important corporate and personal data. This can cause significant damage to the business. Since an attack results in significantly above network traffic than usual, monitoring traffic for abnormal patterns can facilitate the detection of most of the cyber-attacks.

A well designed threat detection system consists of network data collector, streaming module(Kafka, Storm etc.) and Analysis module. Collector will collect the data from network and send the data to Streaming module. Then the stream will be sent to analysis module. Analysis module consists of the latest streaming analysis technologies(Flink, Spark Streaming etc.) and the threat detection mechanisms to detect malicious activities. Cloud environment will give the advantage of concurrently processing huge volume of data and also increase the efficiency of monitoring and threat detection process.

Instead of offline analysis, we can utilize the statistical tools to extract meanings and use these models in the streaming data analysis to detect the threat. Statistical model may become more accurate as volume of data increases and can adapt to changes in the data over time. All these things should happen without compromising the time complexity and efficiency. We can use streaming k-means clustering algorithm and Fuzzy c-means clustering algorithm both of which can identify patterns over time to make the accurate decision [2].

In the below example of threat detection system, a test server is used with ports open for the experimental purpose. The system was trained with normal traffic data with around the size of 260 GB in pcap format containing network traffic packet information.

[Figure 5 about here.]

[Figure 6 about here.]

Figure 5 shows that the port access count is within the range of 100 to 180 per minute. In figure 6, it clearly indicates that there is some abnormal activities happened in between time 220 to 245 [2]. Administrators will be triggered for any such kind of activities immediately.

## 4 IMPROVEMENTS

Having big data in data center operations results improvements in [3],

- Operational Efficiency
- Reduce Infrastructure/Network Down Time
- Improve Customer Experience
- Financial Performance
- Reduce Network/Infrastructure Operations Expenditure
- Reduce Customer Churn

## 5 CONCLUSION

Today's network poses performance and security challenges to network managers because of the layers of redundancy, management and virtualization. The only practical approach for the networking team is to stream or collect all the network related data which is voluminous and varied, into an analytical framework. Big Data analysis is opening up new sales opportunities and risk alleviation

in the networking world. Organizations are already benefiting increased uptime, faster troubleshooting and improvement in security by on-boarding Big Data analytics in the network infrastructure.

## ACKNOWLEDGMENTS

The author would like to thank the web loaded with information. The author would also like to thank Prof. Gregor von Laszewski for his guidance and suggestions.

## REFERENCES

- [1] apmdigest. 2016. Kentik Introduces New NPM Solution. Web page. (September 2016). <http://www.apmdigest.com/kentik-introduces-new-npm-solution>
- [2] Zhijiang Chen, Hanlin Zhang, and William G. Hatcher. 2016. A streaming-based network monitoring and threat detection system. Web page. (June 2016). <http://ieeexplore.ieee.org/document/7516125/>
- [3] dataken. 2017. Datacenter Monitoring and Analytics Platform. web page. (October 2017). <http://www.dataken.net/wp-content/uploads/2015/09/Datacenter-Monitoring-and-Analytics-Platform.pdf>
- [4] Alan R. Earls. 2017. IT analytics tools bring big data to work in the data center. web page. (September 2017). <http://searchitoperations.techtarget.com/feature/IT-analytics-tools-bring-big-data-to-work-in-the-data-center>
- [5] EMC. 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Web page. (April 2014). <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>
- [6] ingrammicroadvisor.com. 2017. How Analytics Can Help Your Customers Improve Data Center Operations. web page. (Feb 2017). <http://www.grammicicroadvisor.com/data-center/how-analytics-can-help-your-customers-improve-data-center-operations>
- [7] ingrammicroadvisor.com. 2017. Key Benefits of Managing Data Center Operations with Analytics. web page. (February 2017). <http://www.grammicicroadvisor.com/data-center/key-benefits-of-managing-data-center-operations-with-analytics>
- [8] David Jacobs. 2016. Network fault management in today's complex data centers. web page. (May 2016). <http://searchnetworking.techtarget.com/tip/Network-fault-management-in-todays-complex-data-centers>
- [9] Hayden James. 2014. 20 Top Server Monitoring & Application Performance Monitoring (APM) Solutions. Web page. (November 2014). <https://haydenjames.io/20-top-server-monitoring-application-performance-monitoring-apm-solutions/>
- [10] Kentik. 2017. Understanding Big Data Network Performance Monitoring: A Tutorial. Web page. (October 2017). <https://www.kentik.com/kentipedia/big-data-network-performance-monitoring/>
- [11] manageengine.com. 2017. Solve your data center management woes with OpManager. Web page. (October 2017). <https://www.manageengine.com/network-monitoring/datacenter-management.html>
- [12] Brien Posey. 2017. How to monitor and manage your data center network. web page. (November 2017). <http://searchnetworking.techtarget.com/tip/How-to-monitor-and-manage-your-data-center-network>
- [13] Murray Slovick. 2017. Volume and Velocity are Driving Advances in Data Center Network Technology. Web page. (September 2017). <https://www.mouser.com/applications/communications-network-technology-advances/>
- [14] Liangwei Zhang. 2016. *Big Data Analytics for Fault Detection and its Application in Maintenance*. Master's thesis. Lulea University of Technology, Sweden. file:///home/dhanya/Downloads/341301390-Big-Data-Analytics-for-Fault-Detection-and-Its-Application-in-Maintenance.pdf

#### LIST OF FIGURES

1	Monitoring and analytical platform [3]	5
2	Total of connected devices (in billions) [13]	6
3	Big data based SaaS NPM solution [10]	6
4	Fault detection architecture [14]	7
5	Ports access count within range 100 to 180 per minute [2]	8
6	Ports access count shows abnormal port activity [2]	9

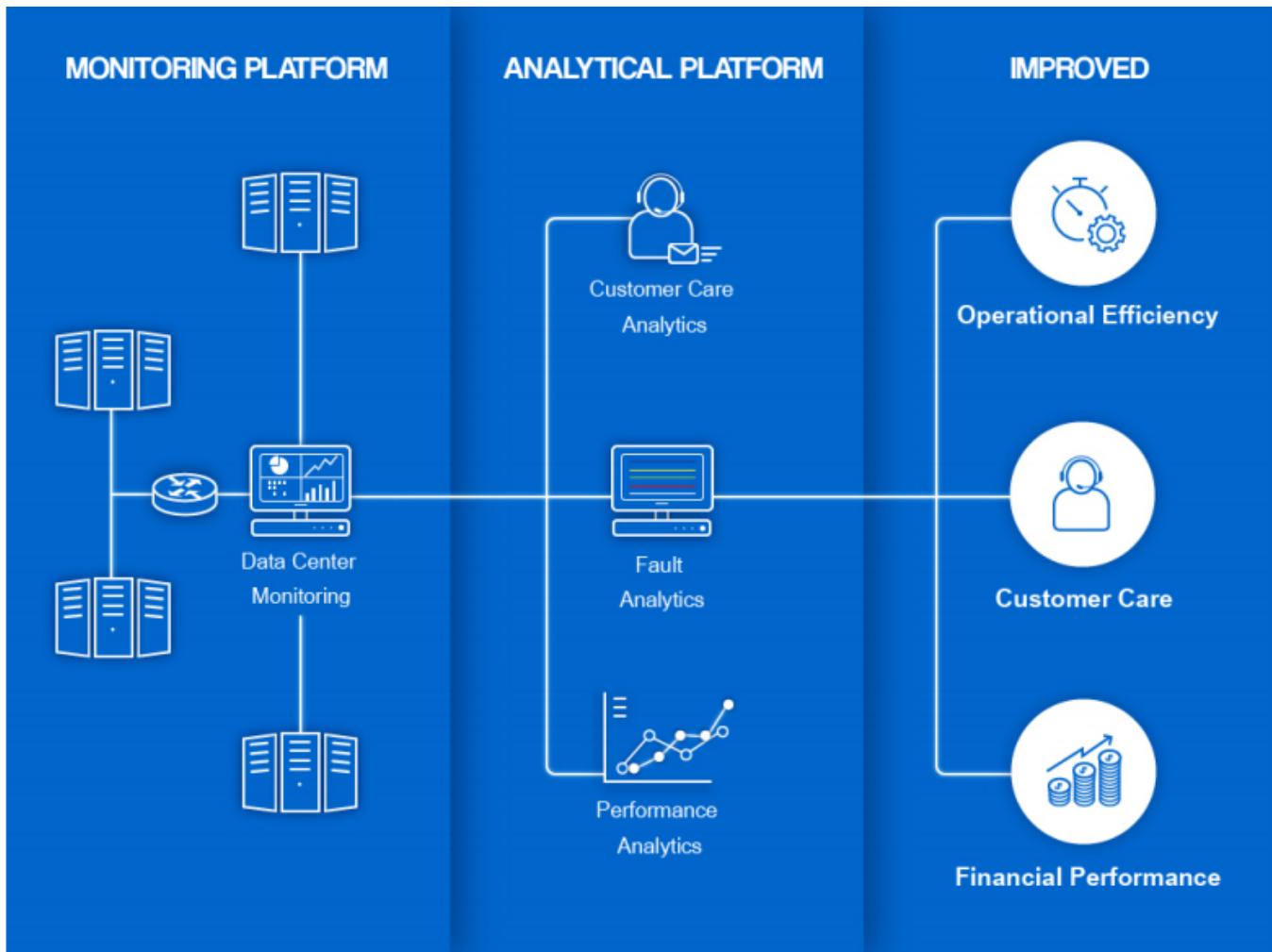


Figure 1: Monitoring and analytical platform [3]

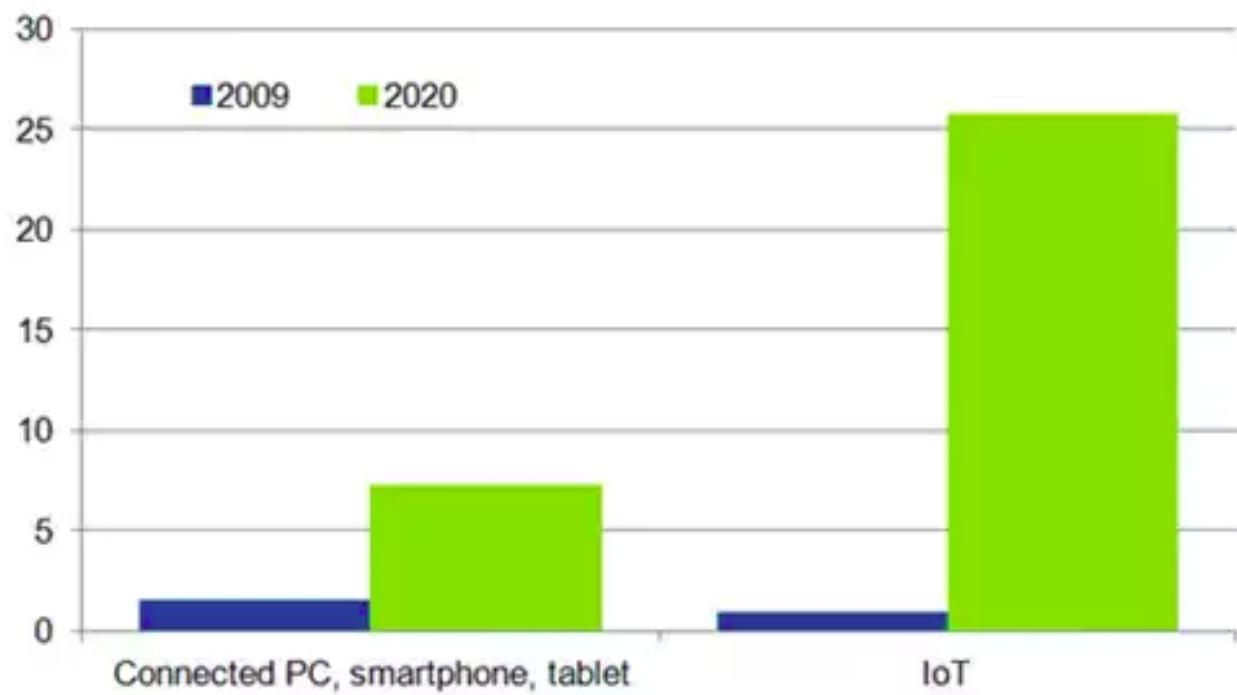


Figure 2: Total of connected devices (in billions) [13]

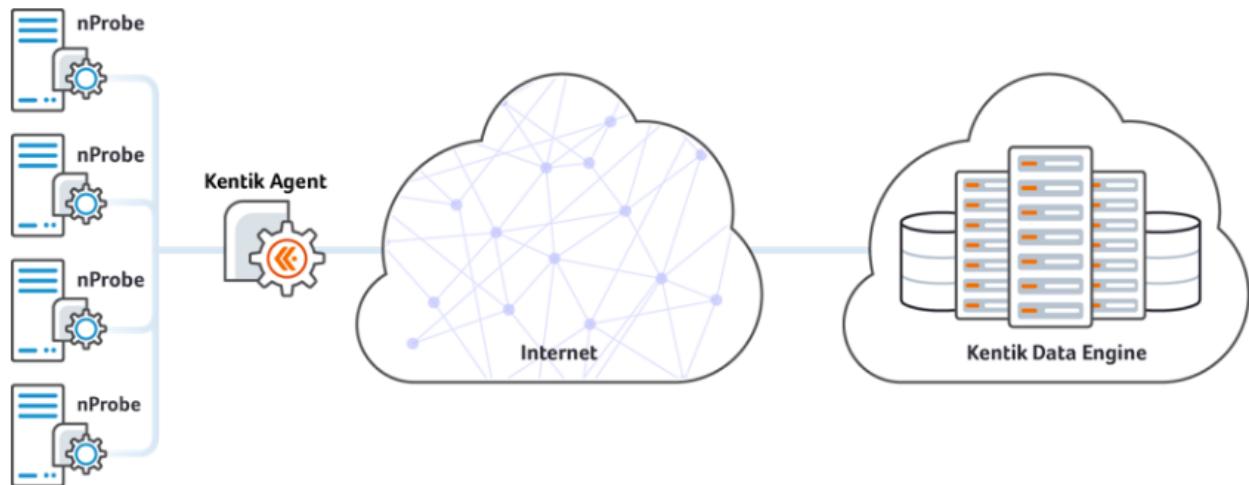


Figure 3: Big data based SaaS NPM solution [10]

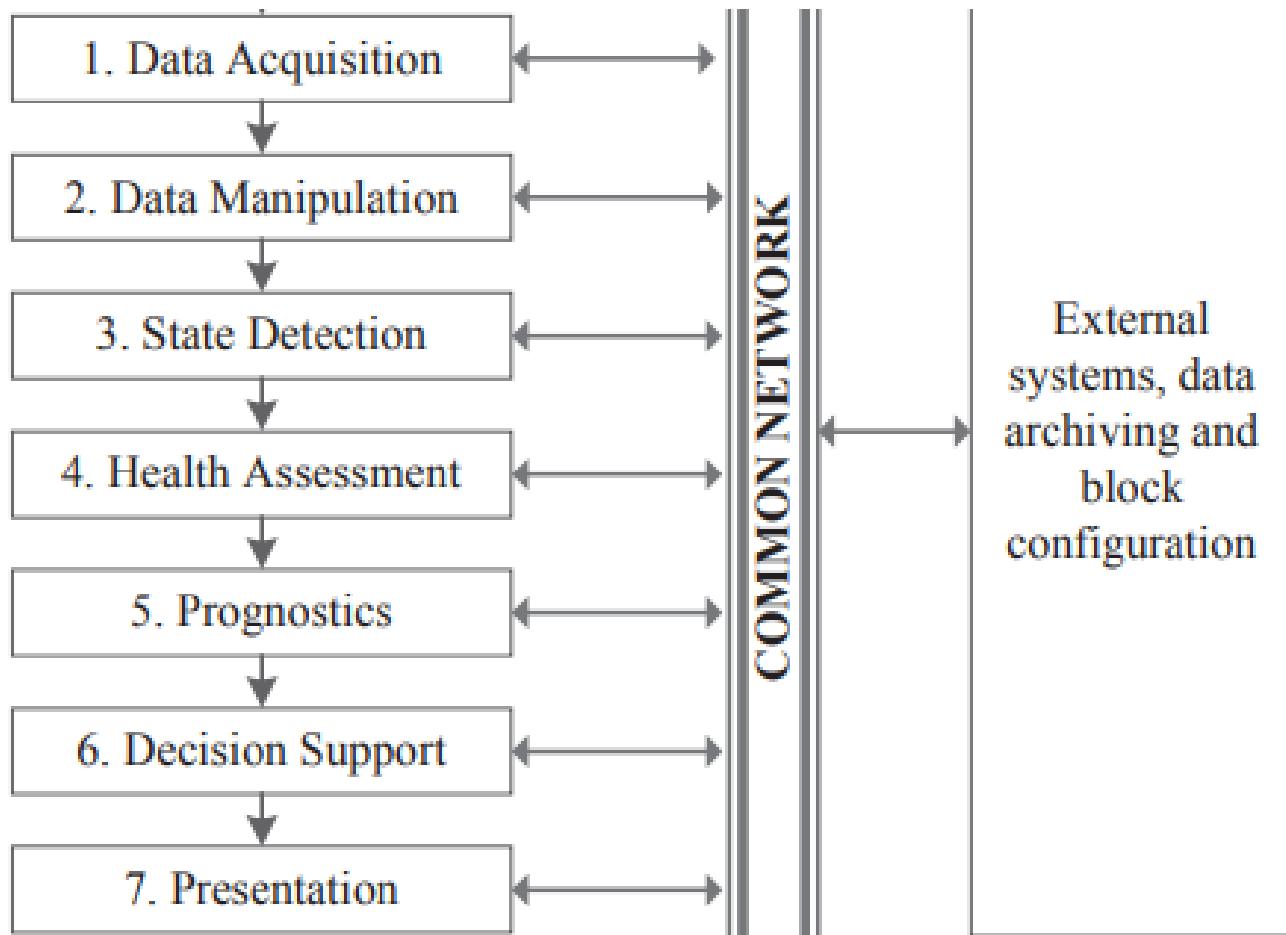
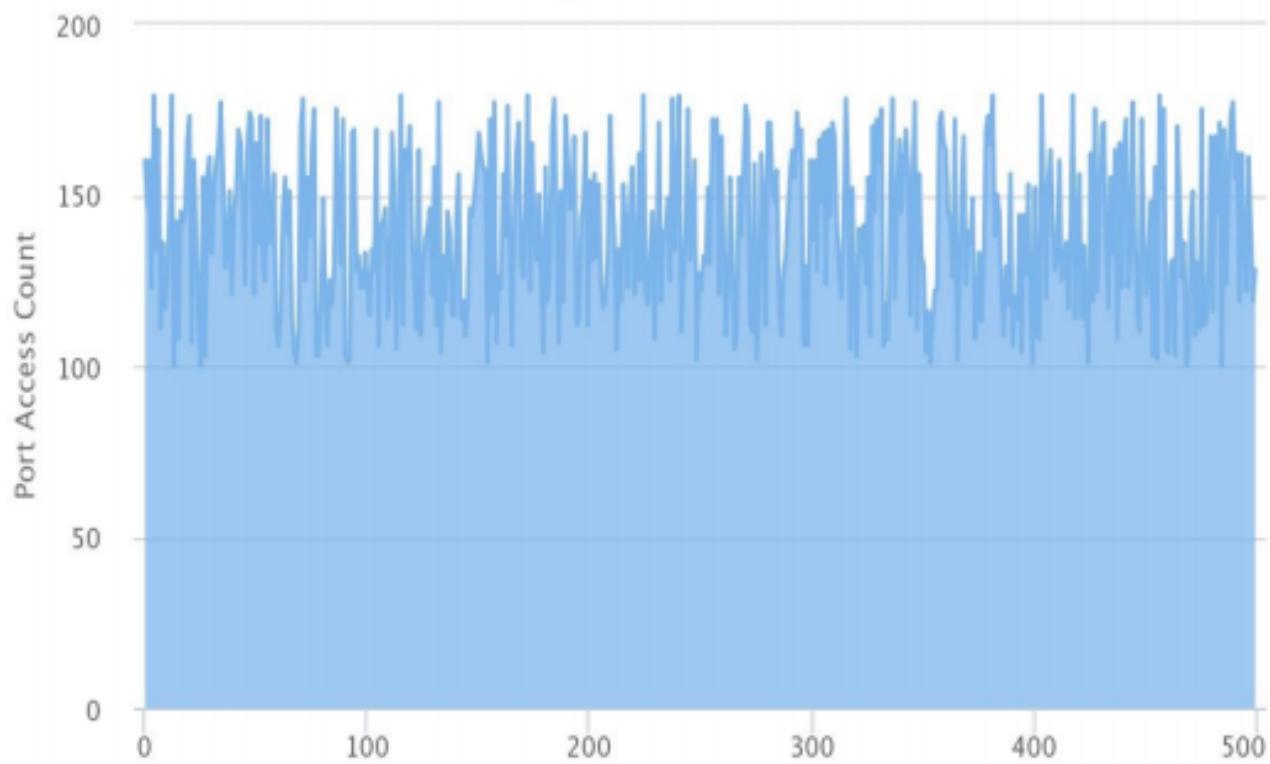


Figure 4: Fault detection architecture [14]



**Figure 5: Ports access count within range 100 to 180 per minute [2]**

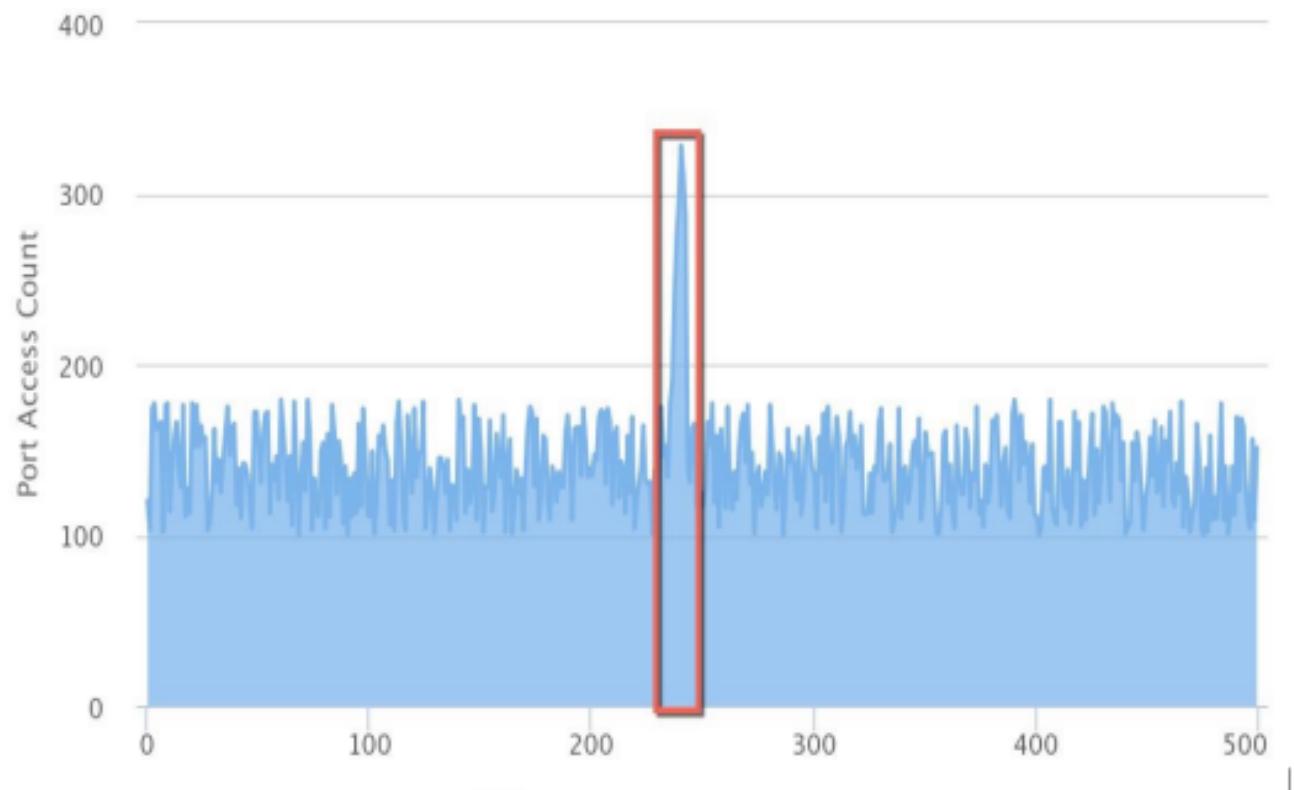


Figure 6: Ports access count shows abnormal port activity [2]

```
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
```

```
[2017-11-06 17.37.48] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Typesetting of "report.tex" completed in 1.3s.
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Dhanya Mathew
hid: 328
paper1: Nov 2 17 100%
paper2: 100%
```

```
yamlcheck
```

```
-----
```

```
wordcount
```

---

```
9  
wc 328 paper2 9 2131 report.tex  
wc 328 paper2 9 2323 report.pdf  
wc 328 paper2 9 485 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
50: \includegraphics[width=1.0\textwidth]{images/Figure1.png}  
53: \label{fig:Figure1}  
56: As shown in Figure \ref{fig:Figure1}, data center monitoring  
platform is capable of monitoring all types of network devices.  
The events captured by the monitoring tools will be passed to the  
big data analytical platform for processing. Here events from  
different monitoring tools will be integrated and related to  
obtain the insights. These insights would be the base for deriving  
decisions and in turn results improvements in operational  
efficiency and financial performance etc.  
94: \begin{figure}[htb]  
96: \includegraphics[width=1.0\columnwidth]{images/Figure2.png}  
99: \label{fig:Figure2}  
102: Figure \ref{fig:Figure2} shows the data generation in billions  
per devices types. This huge data generation by default increases  
the growth of data centers by adding more and more servers and  
network devices. Even a simple addition to the current  
infrastructure would require detailed monitoring in terms of
```

compliance, security and performance.

113: \includegraphics[width=1.0\textwidth]{images/Figure3.png}

116: \label{fig:Figure3}

119: As shown in Figure \ref{fig:Figure3}, nProbe running on host computers send network performance monitoring data securely to Internet via encrypting proxy agent which is optional. Kentik data engine receives this data from internet and stores this data for actionable analytics. This data engine is big data based and can scale horizontally to store unsummarized data and provides powerful analytics for alerting, diagnostics and other use cases.

143: \begin{figure}[htb]

145: \includegraphics[width=1.0\columnwidth]{images/Figure4.png}

148: \label{fig:Figure4}

151: Figure \ref{fig:Figure4}, shows the network fault analysis architecture using big data. The procedure may include below activities as well.

190: \begin{figure}[htb]

192: \includegraphics[width=1.0\columnwidth]{images/Figure5.png}

195: \label{fig:Figure5}

199: \begin{figure}[htb]

201: \includegraphics[width=1.0\columnwidth]{images/Figure6.png}

204: \label{fig:Figure6}

207: Figure \ref{fig:Figure5} shows that the port access count is within the range of 100 to 180 per minute. In figure \ref{fig:Figure6}, it clearly indicates that there is some abnormal activities happened in between time 220 to 245 \cite{streaming-based-network-monitoring-and-threat-detection-system}. Administrators will be triggered for any such kind of activities immediately.

figures 4  
 tables 0  
 includegraphics 6  
 labels 6  
 refs 5  
 floats 4

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

Label/ref check  
 passed: True

When using figures use columnwidth

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

```
find textwidth
```

---

```
50: \includegraphics[width=1.0\textwidth]{images/Figure1.png}
```

```
113: \includegraphics[width=1.0\textwidth]{images/Figure3.png}
```

```
passed: False
```

```
below_check
```

---

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

```
163: Based on bid data classification and clustering algorithms we can
predict below actionable insights.
```

```
bibtex
```

---

```
label errors
```

```
bibtex errors
```

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

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

passed: True

ascii

---

non ascii found 8217

---

The following tests are optional

---

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

find newline

---

passed: True

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

find cite {

---

passed: True

# Big Data Applications in Using Neural Networks for Medical Image Analysis

Tyler Peterson

Indiana University - School of Informatics, Computing, and Engineering

711 N. Park Avenue

Bloomington, Indiana 47408

typeter@iu.edu

## ABSTRACT

Medical image analysis is proving to be a promising domain for disruption by machine learning. The analysis of medical images has long been within the purview of radiologists, a specialization in medicine that reviews medical imaging to form diagnoses and advise on treatment options. Historically, radiologists have relied on their training, senses and years of experience to evaluate images for medical issues, such as the presence of malignant growths, lung nodules, and hip osteoarthritis. The use of technology, generally referred to as computer-aided diagnosis (CAD) tools, has been growing over the last several decades, but modern computing power and sizable datasets have accelerated the effectiveness of these tools. Machine learning algorithms, especially artificial neural networks (ANN), are being leveraged to help identify abnormalities present in medical images at a high level of accuracy. Several research studies conclude that ANN techniques can match, and often greatly improve, the abilities of radiologists. Big data and the application of advanced algorithms show promise for evolving our ability to successfully evaluate medical images and save lives in the process.

## KEYWORDS

i523, hid331, Big Data, Medical Image Analysis, Artificial Neural Networks, Medicine

## 1 INTRODUCTION

The analysis of medical images is primarily the responsibility of radiologists. These individuals are medical doctors who specialize in diagnosing diseases through review of images produced by various imaging modalities, such as x-ray, ultrasound, computerized tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) [8]. Radiologists serve as an expert to other physicians by analyzing the medical images of patients suspected of having certain medical issues, and by making recommendations on subsequent care based on the observations [8]. The images reviewed by radiologists are generally stored digitally, and images are increasingly being stored in picture archiving and communications systems (PACS). These systems are required to keep up with the rapid accumulation of medical image data. Between 2005 and 2011, the medical image data in US hospitals increased from only 8,900 terabytes to 27,000 terabytes [9]. That number is expected to grow by 20 percent every year due to increasing image size and resolution, the adoption of 3D imaging, and an aging population who will likely bring an increasing demand for medical imaging studies [9].

It is estimated that one billion medical images are created worldwide each year, and most of these are assessed by radiologists [2]. Given that radiologists are human, their judgment is fallible. It is estimated that the lowest average error rate in analyzing medical images is 4 percent, which means collectively radiologists are estimated to make 40 million errors in judgement every year [2]. A particularly striking example of fallibility comes from a study that analyzed the first and second interpretations of radiologists from Massachusetts General Hospital. They reviewed abdominal CTs and re-reviewed studies that had either been interpreted by themselves or a colleague. The study found that the radiologists disagreed with their peers 30 percent of the time and even disagreed with themselves 25 percent of the time [2].

There are two major types of radiologic analysis error: perceptual error and interpretive error [2]. Most errors, up to 80 percent, are perceptual errors, which occur when an abnormality is not perceived by the reviewer during the initial review, but is identified in a subsequent analysis [2]. Interpretive errors occur when the radiologist successfully identifies the abnormality, but incorrectly diagnoses the problem, which may lead to a less appropriate course of care [2]. There are several reasons why errors occur, including fatigue, excessive pace of analysis, distractions and insufficient knowledge of the practitioner. It is also asserted that the extreme complexity of a radiologist's job contributes to the errors. Errors occur in the practice of radiologists all across the world, at varying levels of training, in all imaging modalities and all clinical settings [2].

Over the last several decades, there has been an effort to develop computer-based tools to aid radiologists in the detection of abnormalities. Computer-aided diagnosis (CAD) tools are primarily intended to increase the rate at which problems are identified while also reducing the false negatives resulting from human error [12]. These systems are intended to supplement, not replace, the radiologist by reporting a second opinion to be considered alongside the radiologist's assessment. The earliest initiatives to develop these tools occurred in the 1960s, and concerted efforts began in the 1980s [3]. Despite the research being nearly 60 years old, widespread adoption is a relatively recent occurrence [5]. Clinical studies reported early CAD implementations as being minimally effective. Specifically, CAD decisions included more false positives than human assessments, which created more work for radiologists and often led to additional, unnecessary medical tests and biopsies [6].

Several improvements in the field of computing have increased the accuracy of CAD tools and subsequently encouraged wider adoption of these tools into clinical workflows. The advancements

includes increased access to digital imaging datasets, larger imaging datasets, increased used of imaging in healthcare and increased computing power [5][6]. These factors combine to create an ideal state for research related to advanced machine learning techniques, namely artifical neural networks (ANN), and the implementation of tools that can rival the assessment of highly trained radiologists.

## 2 MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORKS

Broadly speaking, machine learning is a way of applying artificial intelligence to a problem through the analysis of data. Machine learning techniques assess the features, or attributes, of samples in a dataset to identify patterns in the data, and the resulting algorithm can be used to render conclusions about new inputs without human intervention [6]. The ideal algorithm is represented by an equation that minimizes the error, or cost, made by the predictions. Medical image analysis presents what is referred to as a classification problem. The typical example of a classifcation problem is handwritten numerical digit recognition. In this example, a handwritten digit between 0 and 9 is fed into the algorithm, and the algorithm decides which of the ten digits, or classes, that handwritten digit is most likely to belong. Specific to medical image analysis, the classifier detects abnormalities in images otherwise not present in images of the same area in healthy individuals, and renders a conclusion as to what that abnormality is.

There are several different machine learning techniques that have traditionally been used in classification problems, such as support vector machines (SVM). SVMs are an example of a supervised machine learning model, meaning this method uses labeled data. Every sample in the dataset includes the label, or correct answer, along with a value for each of the attributes. The SVM identifies patterns in the labeled samples to create an algorithm, and new, unlabeled samples can be processed by the algorithm and given a prediction. In the task of medical image analysis, attributes have historically been identified and designed by human experts [12]. For example, an expert would identify abnormalities by explicitly describing shape, texture, position and orientation of the abnormal biological structure [10]. In addition to this being labor intensive, the image features are specific to the immediate problem being explored and cannot be expected to work well for other image types [12].

ANNs are another class of machine learning that is increasingly being leveraged to tackle problems related to image analysis. ANNs, just like SVMs, are often utilized in a supervised manner, but do not require the painstaking process of expert-defined key attributes. Instead, ANNs learn the important features from the images themselves [10]. This is an obvious advantage when compared to traditional machine learning methods, but the complexity of these algorithms has prevented the achievement of mainstream use until recently. The theory of ANNs was introduced in the 1950s, and research in this field has begun to flourish recently due to the increase in computing power and availability of high quality datasets [6].

## 3 OVERVIEW OF ANNS

The inspiration for the design of ANNs is the biological neural network, more commonly known as the brain. The process by which data is input into the model, analyzed, and given an output is intended to mimic the way a brain absorbs and processes information before finally coming to a conclusion. In the brain, each neuron is capable of receiving multiple input signals and transmitting those signals via synapses to other neurons [6]. Similar to the brain, ANNs are made up of artificial neurons that take in inputs, or attributes, from the samples of the dataset. In the context of medical imaging, the inputs are numerical representations of each individual pixel of the original image. And just as neurons in the brain transmit information to other neurons via synapses, the artificial neurons pass information via artificial synapses. Each time information travels by way of artificial synapses, that information is multiplied by a weight [6]. As with traditional machine learning techniques, the weights are intended to minimize the error, or cost, of the function, thereby returning a higher accuracy rate. The learning process of the ANN is driven by the adjustment of those weights, similar to how the neurons in the brain use external stimuli to adjust and redistribute evaluations. The weights in the ANN algorithm are initially randomized and subsequently adjusted by an optimization algorithm such as gradient decent, which guides the weights in a direction that minimize the cost of the function [6].

There are several types of ANNs, and the types can be identified by the structure. The most basic neural network, called the perceptron, was initially theorized in 1957 and consisted only of an input layer and an outputer layer. This design limited the perceptron's problem solving ability to datasets that could be linearly separated [15]. This is clearly not helpful for problems such as medical image analysis where the data presents complicated patterns and relationships.

In 1982, the Hopfield network was theorized, which adds a hidden layer of artificial neurons between the input and output layers [15]. Its referred to as a hidden layer because the values of the neurons in the hidden layer do not correspond to a specific input value or a output class prediction [6]. Each input neuron is connected to each neuron in the hidden layer, and each neuron in the hidden layer is also connected to each neuron in the output layer. It is important to note that while all neurons in neighboring layers are connected to each other, the neurons within a single layer are not connected to each other [12]. The benefit of the additional layers is that it allows the neural networks to combine numerous simple decisions to make more complicated decisions [6]. Deep neural networks (DNN) are type of ANN that make use of several hidden layers between the input and output layers, and have demonstrated the capability of making complex decisions.

Convolutional neural networks (CNN) are an advanced type of ANN that are well suited to solving problems related to images. The fact that neighboring pixels are directly next to each other or near each other is an important piece of information that can tend to be lost by other types of ANNs that vectorize input values. CNNs, on the other hand, input images in a more direct and complete manner [12]. CNNs are comprised of several types of layers, including convolutional, pooling and fully connected layers. Convolutional

layers detect distinctive edges, lines and other perceptible visual features. This is intended to mimic how the brain perceives objects by observing distinct visual features [6]. Pooling layers get that name because they pool together the image in a way that reduces the dimensions of the input sample while preserving the important details identified in the convolutional layer. Convolutional and pooling layers are often repeated several times before arriving at a fully connected layer, which ingregrates the results from the previous convolutional and pooling layers [6].

Several statistics can be used for evaluating the accuracy of ANNs. Sensitivity is the true positive detection rate. This is the percentage of positive occurrences that are successfully identified [4]. A false positive may lead to unnecessary testing, unnecessary expenses and unnecessary stress on the patient who has been led to believe they have a certain condition. Specificity is the true negative detection rate. This is the percentage of negative occurrences that are successfully identified [4]. A false negative may lead to a missed diagnosis, resulting in delayed treatment and a potentially avoidable death in some cases. Sensitivity and specificity can be assessed together by the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate (100 minus the true negative rate) for varying decision thresholds. This illustrates the trade-off between sensitivity and specificity and can provide guidance on which decision threshold is appropriate for the task [4]. ROC curves are often leveraged to evaluate the performance of ANNs 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 [4].

## 4 APPLICATIONS IN MEDICAL IMAGE ANALYSIS

There are several specialities in which medical image analysis has been studied and applied for the purpose of computer-aided detection and diagnosis, and the effectiveness of ANNs has been formally studied in different ways. This includes ANNs versus medical professionals, ANNs combined with medical professionals versus ANNs and medical professionals separately, and one type of ANN versus another type of ANN.

One study applied deep learning techniques to images of breast sentinel lymph nodes and evaluated the images for the presence of metastasis. A positive test likely means the staging of the breast cancer will be higher and subsequent treatment will be more aggressive. A false negative means a patient's disease will be thought of as less advanced than it actually is, and subsequent treatment will not be as aggressive as necessary. The medical professionals in this study obtained an AUC of 0.966, and the algorithm received an AUC of 0.925. Decisions made by a human also using the algorithmic conclusion as a second opinion achieved an AUC of 0.995, which equals an 85 percent reduction in error for the medical professional [14].

A second study compared the performance of three different types of ANNs on the detection of cancerous lung nodules in CT images. Lung cancer is a disease that greatly benefits from early detection. Over 220,000 new cases were identified in 2015, and an early detection of the disease improves the 5 year survival rate

by roughly 50 percent [13]. CT images provide three-dimensional (3D) views of the chest and are a key component of the clinical workflow of this specialty. These image are analyzed to understand if the structures in the image are part of the expected anatomy or if the structure is a tumor. If a tumor is present, the goal is to understand if the nodule is benign or malignant. This determination closely depends on the size, shape and texture of the nodule, all of which can be analyzed by an ANN. This study compared the performance of a DNN, a CNN and another type of ANN called a stacked auto-encoder (SAE). The CNN performed best with an AUC of 0.916, while the DNN and SAE recorded AUCs of 0.877 and 0.884 respectively [13].

A third study compares the performance of an ANN to two experienced physicians in the evaluation of x-rays for the presence of hip osteoarthritis. Hip osteoarthritis causes pain and stiffness which can diminish quality of life through an inability to perform daily tasks or go to work [16]. Hip osteoarthritis is diagnosed through x-ray imaging studies, which are traditionally evaluated by radiologists. The time consuming and error-prone nature of this work is well documented, and with an increasingly aging population, efficient and timely diagnosis of hip osteoarthritis is growing in importance. The study used a CNN to achieve a sensitivity rate of 95.0 percent and a specificity rate of 90.7 percent, with an AUC of 0.94. Both physicians achieved a sensitivity of 100 percent and specificity rates of 86.0 percent and 93.0 percent. While the CNN recorded lower sensitivty compared to the two physicians, it did record a higher specificity than one of the physicians [16]. This shows promise for the performance of ANNs in evaluating hip osteoarthritis.

## 5 INFRASTRUCTURE

Optimization of an ANN can require billions of calculations, if not more, and the task therefore requires special hardware to help accomplish the task in a reasonable time frame. At a very high level, there are two tasks involved in training a model. First, the input data is passed forward through the neurons which provides an output and an accompanying error rate. Second, with the error rate in hand, the weights of each synapse in the model are adjusted with the goal of lowering the error rate. This process is repeated many, many times. A common deep learning deployment called VGG16 has 16 hidden layers and roughly 140 million parameters [11]. At each point in the network the computer must complete a matrix multiplication task, and the sheer volume of calculations means the task will take a significant amount of time to complete [11][1].

Graphical processing units (GPUs) are better suited for this task than central processing units (CPUs). The key difference between GPUs and CPUs is that GPUs can parallelize the matrix operations necessary to train the model, whereas CPUs are far less able to do so. CPUs typically only have a handful of cores, while GPUs can contain hundreds, if not thousands [1]. GPUs can perform several matrix operations at once and CPUs need to perform those same operations one at a time. For example, training a CNN with four hidden layers takes 8,000 seconds with a CPU and only 1,000 seconds with a GPU [1].

This foundation has led to the creation of GPU-based super computers. The Commonwealth Scientific and Industrial Research Organization (CSIRO) acquired a new super computer made by Dell in 2017. Part of the infrastructure includes 114 PowerEdge C4130 server with Nvidia Tesla P100 GPUs, which includes over a million computing cores and 29TB of RAM [7].

## 6 CHALLENGES

There are several challenges that may inhibit the widespread use of CAD tools built upon ANNs. First, overfitting occurs when an algorithm is trained based on a dataset that does not generalize well to examples outside of the data used to train the algorithm. Many studies demonstrating the value of ANNs were trained using relatively small datasets, and because the significant features present in a small dataset may not be the same features present in a large dataset, the algorithm derived from the small dataset may not perform well when analyzing images from the large dataset [15][6]. This issue can be addressed by training algorithms on larger datasets, but of course requires access to larger datasets, longer training periods and more computing power.

Second, algorithms derived from ANNs are frequently considered to be black boxes, meaning that it is nearly impossible to understand how the algorithm reached a certain conclusion. This contrasts with several other types of machine learning techniques that produce equations that highlight which features are significant [6]. Instilling belief and trust in a system that is difficult, if not impossible, to explain is a barrier, even if that system produces accurate responses the vast majority of time.

Third, ethical and legal considerations must be made. Adopters of this technology must consider scenarios where the system makes a prediction that harms a patient [6]. If a radiologist is led to a conclusion by an algorithm, and the algorithm presents a false positive, a false negative or presents one conclusion while missing another, who is responsible for the error?

Fourth, the ANNs are dependent on the quality and nature of the imaging data used to train algorithms. There is variability around the world in regards to the type and quality of imaging machines and the imaging protocols that dictate why and how images are taken [6]. Two different imaging machines taking a picture of the same body site may produce meaningfully different images. Further, two technicians may use the same machine differently when imaging the same body site, and this may also produce meaningfully different images. It is conceivable that these images could appear different to an algorithm to the extent that the prediction is not the same. This issue could at least partially be addressed by sufficiently large datasets containing labeled images of abnormalities that were taken using machines of varying quality and executed using differing methods.

## 7 CONCLUSION

ANNs represent a great step forward in our ability to program computers to rapidly evaluate information in a manner similar to that of human experts. The process demonstrates great capabilities in learning important features programmatically, as opposed to researchers needing to consult with experts to handcraft meaningful features. Widespread adoption of this technology is beginning to

pick up speed as the accuracy of these algorithms approaches that of experts. While research does not yet conclusively indicate that algorithms can independently outperform experts, at the very least the combination of an expert and a modern CAD tool frequently leads to higher accuracy compared to the expert operating alone. Continued advances in computing technology and the accumulation of larger and larger imaging datasets will likely further increase the power of these tools.

## ACKNOWLEDGMENTS

Thank you to Gregor von Laszewski and his teaching assistants for their help with the class's numerous questions and concerns.

## REFERENCES

- [1] Dimitrios Bizopoulos. 2017. GPU vs CPU in Convolutional Neural Networks using TensorFlow. Online. (2017). <https://relinklabs.com/gpu-vs-cpu-in-convolutional-neural-networks-using-tensorflow>
- [2] Michael Bruno, Eric Walker, and Hani Abujudeh. 2015. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *RadioGraphics* 35, 6 (2015), 1668–1676. <https://doi.org/10.1148/rg.2015150023>
- [3] Kunio Doi. 2007. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Comput Med Imaging Graph* 31, 4-5 (2007), 198–211.
- [4] 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.
- [5] Tae-Yun Kim, Jaebum Son, and Kwang-Gi Kim. 2011. The Recent Progress in Quantitative Medical Image Analysis for Computer Aided Diagnosis Systems. *Healthcare Informatics Research* 17, 3 (September 2011), 143–149. <https://doi.org/10.4258/hir.2011.17.3.143>
- [6] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk-Bae Kim, Joon Beom Seo, and Namkug Kim. 2017. Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology* 18, 4 (2017), 570–584. <https://doi.org/10.3348/kjr.2017.18.4.570>
- [7] Asha McLean. 2017. CSIRO receives deep learning supercomputer from Dell EMC. Online. (July 2017). <http://www.zdnet.com/article/csiro-receives-deep-learning-supercomputer-from-dell-emc/>
- [8] Radiological Society of North America. 2017. What Does a Radiologist Do? Online. (April 2017). <https://www.radiologyinfo.org/en/info.cfm?pg=article-your-radiologist>
- [9] Morris Panner. 2015. What's Next for the Healthcare Data Center? Online. (April 2015). <http://www.datacenterjournal.com/whats-healthcare-data-center/>
- [10] Faizan Shaikh. 2017. Deep Learning vs. Machine Learning - the essential differences you need to know. Online. (April 2017). <https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/>
- [11] Faizan Shaikh. May. Why are GPUs necessary for training Deep Learning models? Online. (2017 May). <https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/>
- [12] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering* 19 (June 2017), 221–248. <https://doi.org/10.1146/annurev-bioeng-071516-044442>
- [13] QingZeng Song, Lei Zhao, and XingKe Luo and XueChen Dou. 2017. Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images. *Journal of Healthcare Engineering* 2017 (August 2017), 1–7.
- [14] Dayong Wang, Aditya Khosla, Rishab Gargya, Humayun Irshad, and Andrew H Beck. 2016. Deep Learning for Identifying Metastatic Breast Cancer. Online. (June 2016).
- [15] Shijun Wang and Ronald M. Summers. 2012. Machine Learning and Radiology. *Medical Image Analysis* 16, 5 (July 2012), 933–951. <https://doi.org/10.1016/j.media.2012.02.005>
- [16] Yanping Xue, Rongguo Zhang, Yufeng Deng2, Kuan Chen, and Tao Jiang. 2017. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLoS ONE* 12, 6 (June 2017), 1–10. <https://doi.org/10.1371/journal.pone.0178992>

## 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 inflate 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 float. 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

```
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
```

```
[2017-11-06 17.37.57] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Label(s) may have changed. Rerun to get cross-references right.
Typesetting of "report.tex" completed in 1.5s.
```

```
=====
```

```
Compliance Report
```

```
=====
```

```
name: Tyler Peterson
hid: 331
paper1: Oct 22 17 100%
paper2: Nov 6 17 100%
project: Dec 04 17 0%
```

```
yamlcheck
```

---

```
wordcount
```

---

```
6
```

```
wc 331 paper2 6 3605 report.tex  
wc 331 paper2 6 4470 report.pdf  
wc 331 paper2 6 740 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

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

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

```
Label/ref check
```

passed: True

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

ascii

---

```
=====
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

# Advancements in Drone Technology for the U.S. Military

Peter Russell  
Indiana University  
petrusse@iu.edu

## ABSTRACT

Technological breakthroughs in military technology have put the U.S. in a new chapter of warfare. These advancements have become realizations of what was at one time only deemed possible in science fiction, such as autonomous decision making and weaponization of drones. These innovations provide unique advantages to the leaders of the technology and are too lucrative to ignore. However, in coming years as these technologies push the boundaries, decisions will need to be made in how much control military leaders are willing to give to their new mechanic allies and whether they should be passive, as they have been in the past, or active participants on the battlefield.

## KEYWORDS

i523, HID 334, Drone Technology, Big Data, Military Technology

## 1 INTRODUCTION

Among the many industries being transformed by the Big Data movement, few carry more consequences than the changes being experienced in the U.S. military due to the direct impact on human life. It has been argued that the current changes could affect warfare and diplomatic landscape on the same scale that nuclear weapons did [1].

Traditionally, drones can be categorized as a re-usable, autonomous vehicle either in the air or on the ground. In the air, these are known as “Unmanned Aerial Vehicles”, or UAVs, and on the ground, “Unmanned Ground Vehicles”, or UGVs. More recently, this has expanded to USVs for “Unmanned Surface Vehicles” and UUVs for “Unmanned Underwater Vehicles.” Our focus will remain primarily on the former two types, UAVs and UGVs since these have been the most long-standing types.

For most citizens, UAVs remain the more well known of the two and have garnered more attention recently for their proposed commercial and consumer uses. Unsurprisingly, this has created rapid growth in the industry. In 2017 alone, the \$6 billion industry for these uses is expected to grow by 35%, roughly matching its 2016 growth [9]. While these market segments are growing quickly, they remain in their nascent stages and dwarfed by drone spending in the Department of Defense (DoD). To put the difference of size in perspective, for the FY2018 budget, the DoD requested nearly \$7 billion *for the year* in drone spending, which is the highest since FY2013 and \$3.3 billion than previously estimated four years ago [10]. This annual spending is no anomaly. Currently, the DoD is responsible for nearly 90% of the spending in the UAV market [7].

This continued investment in the drone program demonstrates a clear vote of confidence in the advancement of this technology and its impact on military operations.

## 2 STRATEGIC ADVANTAGE WITH DRONES

One facet in the complexity of military planning is finding the path that gives the highest probability of success with the lowest possible risk of casualties. In this light, the stakes of the problem that technology is trying to solve with Big Data could not be higher. Leaders are constantly trying to solve a constrained optimization problem and when it comes to this overarching problem of mission success, Big Data can be utilized to help the decision maker solve the smaller sub-problems that comprise it, such as where to set up surveillance, where and when to pursue the enemy, how to carry out reconnaissance and how to disarm hazardous obstacles along the way.

### 2.1 Surveillance

The RQ-4 Global Hawk is currently America’s most expensive surveillance drone, projected to cost nearly \$428 million in 2018 [10]. Having flown missions for nearly 15 years now, the total cost of this drone is over \$14 billion [6]. This drone provides a great case study in the evolution of military drone capabilities for its long track record, which stands at over 200,000 flight hours [11].

Initially, the RQ-4 provided imagery through its equipped sensors, which were for landscape topography (synthetic aperture radar), navigation (electro-optical sensors) and heat signatures (infrared sensors). Later models have been equipped with features that allow the antennae to move on its own for an improved signal and a radar tracking system that allows its operators to zero in a target from its surroundings (moving target indication).

The advancement of sensors provides leaders with high resolution photos for strategic planning. In its current capability, images can be obtained with up to a 1 foot resolution and targeting precision within a 60 foot radius from its maximum height of 65,000 feet [21]. However, it should be noted, the revolutionary aspect in drones does not necessarily come entirely from its sensors. The U-2, which was first introduced in 1955 and famously downed over the Russian border during the Cold War, can be retrofitted with these sensors. The stark difference modern aerial surveillance has with its predecessors like the U-2 is that the current technology allows the device to be unmanned. As a result, if a situation were to occur like with the U-2 where the plane is downed, there is no pilot to capture and a diplomatic crisis around hostage negotiation is avoided. Along these lines, a similar situation played out in late 2016 when China captured an underwater U.S. drone with no major diplomatic ramifications. Additionally, unmanned drones allow for flight times that would likely push beyond the boundaries of human focus and endurance since the RQ-4 can sustain flights in excess of 30 hours [19].

## 2.2 Swarms

One of the most exciting applications of drone technology revolves around drone swarms. Spending in this category, broadly defined as “Autonomy, Teaming and Swarms” has doubled in the last four years [10]. With UGV spending stagnant over this period, this program is now receiving twice as much funding, but is still only a small fraction of the largest program, unmanned aircrafts, at 10% of that spending. The rapid growth in this program will undoubtedly continue given the revolutionary nature of these swarms as it relates to warfare.

The public has been aware of this new technology since early 2016, but its development has been underway since at least 2014 [13]. The program, however, did not gain mass attention until a 60 Minutes special aired in early 2017 introducing the Perdix drone and demonstrating a mock swarm mission comprised of 103 Perdix drones acting as a single unit [15].

The Perdix drone looks similar to a toy airplane, weighing only a pound and with a wingspan of 6.5 inches. This simple design reflects the expendability of each drone, which is one of the swarms major advantages. In the swarm, there is no lead drone in the swarm and therefore, no single vulnerability to attack if one of the drones was taken down by the enemy [5]. As a result, each drone is designed to work with other drones of the same type as a single unit to achieve a given mission objective and fill in any gaps if drones are no longer functional. These drone swarms are intended to be able to scan large areas very quickly, provide electronic jamming against the enemy, create a wide communication area for ground troops or confuse enemy radar [16].

In the 60 Minutes demonstration, these Perdix drones were dropped from F-18 jets at the speed of sound, aggregated together and collectively scanned an area, entirely on their own. The innovation in computing and Big Data allows the swarm to exist since no human either individually or as part of a team could make the calculations that these drones are making collectively to achieve their mission.

At the moment, while also being a means of surveillance like the RQ-4, these swarms are not a replacement for these traditional drones, nor does that seem to be the end goal. These Perdix drones have a flight time of only 20 minutes currently and are flown at a relatively low-altitude. This compares with the RQ-4, which is considered a HALE, or High Altitude Long Endurance drone. Additionally, the RQ-4 requires a team of nearly a dozen while the swarm is given a directive by an operator on its objective and requires no human intervention [8]. Lastly, with a unit cost of \$235 million per unit, the RQ-4 holds an economic liability with any enemy attack that the Perdix does not at only \$30,000 per unit.

Eventually, these swarm drones are expected to have the capability to be aggregated together by the thousands and carry out overwhelming and confusing attacks on enemies. It has been properly described as the “difference between a wolf pack and just little wolves[8].”

## 2.3 Disarmament and Detection

Of these two drone segments, UAVs remain by far the larger of the two with spending on UAVs nearly 20x that of UGVs[10]. To date, UGVs have been responsible for aiding ground troops in their

mission. While this could come in the form of reconnaissance or in helping carry heavy loads, explosive detection has arguably been the most important impact for their ability to screen areas for improvised explosive devices (IEDs) along paths that ground troops must travel to complete their mission.

In comparison to aerial innovations, UGV development has developed at a slower pace when it comes to full autonomy. This is largely due to the nature of challenges a ground drone faces in navigation versus flying. Namely, how to deal with uneven terrain and unpredictable obstacles [14]. Nonetheless, user operated UGVs have proven to be a tremendous advantage as it relates to disarmament and detection.

To circumvent the endless and unique possible situations a UGV could be faced with, the military been innovative in the way these UGVs are deployed instead to avoid these hurdles. For example, soldiers can now throw a five pound UGV from a height of up to 15 feet to begin a reconnaissance or bomb detection mission [18]. This allows them to be thrown on top of a roof or into openings that humans might not be able to fit. These robots are equipped with video cameras and various sensors to relay information about the landscape back to the operator.

## 3 RECENT DEVELOPMENTS

In the field of surveillance, one of the newest drones being pursued is the Zephyr 8, a solar powered drone that can fly for 45 days continuously. This flight time allows the drone to be launched in the U.S. and reach destinations like Afghanistan on its own, but perhaps even more incredible is the amount of data this drone can produce. Specifically, it flies at a height of nearly 12.5 miles in the sky, far exceeding the height needed to see the curvature of the Earth, but can still take pictures at the precision of 6-inch resolution. This height allows surveillance of 386 square miles and coupled with this resolution, this becomes a large data set very quickly [3]. One of the newest developments in drone technology by the U.S. military does not categorize as a UAV or UGV, but instead as a USV, for Unmanned Surface Vehicle. These are autonomous boats with the most famous example to date being the Sea Hunter, which was introduced in 2016. This massive vessel, with the length of 132 feet and 135 tons, was built to track diesel submarines and detect mines [17]. It is a major innovation for the U.S. military for its range, which is 12,000 miles on a single tank of gas, and its economic savings, which is 2% of what a traditional ship costs to operate daily. [20] [4]. Or, framed differently, the U.S. military can operate 50 of these Sea Hunter ships for the same cost as one traditional ship. This has proven to be a Big Data and computational marvel as the ship operates autonomously through 36 computers running 50 million lines of code [15].

## 4 FUTURE DEVELOPMENTS

One of the aspirational areas of future drone development for the military is in the field of Micro Air Vehicles (MAVs), which as the name implies, are extremely small UAVs, such as the size of a small bird. Even within that area, there is a growing interest in Nano Air Vehicles, which could be the size of an insect. The future of this technology is for troops to gain intelligence on areas that would be either too dangerous or physically impossible to enter.

One of the more well-known MAVs is the Black Hornet Nano. The drone measures 4 inches in length and is an inch wide with the weight of just a half an ounce, or the weight of 3 pieces of paper. This drone has three cameras and can fly for 20 minutes non-stop. Interestingly, the drone is designed to stream video back to its operators to avoid the risk of footage being compromised if it were stored locally. While this is all extremely impressive, future developments are pushing to make these MAVs even smaller. However, the smaller and lighter these MAVs become, the harder they become for a user to control. The reason being is that the smaller they are, the more sensitive they become to cross winds, the more difficult they are to equip with navigation sensors and the smaller field of vision the camera has. However, the inability to be detected by enemies is a tremendous advantage and to circumvent these piloting issues, work is being done to make them fully autonomous, potentially even as a swarm.

## 5 INTEGRATION OF DRONE TECHNOLOGIES

One of the beautiful aspects of technological innovation are the synergies created. For the U.S. military, these synergies in the context of the Big Data movement are opening new possibilities with difficult questions that will eventually have to be answered. An example of this is how or if drones should be weaponized, even if their decision making is superior to humans.

Without Big Data this debate could never take place. For example, one of the highest resolution drone surveillance cameras in 2014, ARGUS-IS, disclosed some, but not all, of its features as some parts remained classified. It was equipped with a 1.8 billion megapixel camera that could monitor 10 square miles and store all of this information, which works out to be 6 petabytes of data daily [2].

This information accumulation allows greater monitoring of potential targets. Namely, if a known target is tagged and tracked, pictures can be taken at different angles and stored in a database. This ability to accumulate a massive amount of data improves the accuracy of facial recognition. One demonstration showed how a low-altitude drone could be coupled with a UGV and USV against an enemy. The low-altitude and UGV would work in conjunction with each other to carry out a reconnaissance and scan an area and once a match of the target has been found, communicate this to the USV in a different location to fire the weapon systems on the target [15].

While this chain of events is currently possible, which is to attack an enemy with no human interaction, there is a difficult ethical choice to be made in how, or if, these drones will be integrated with respect to weaponization. Even if computers are able to make better decisions on facial recognition, which recent evidence suggests that they can, there remains a large reluctance to open this potential Pandora's box as a new type of warfare [12].

## 6 CONCLUSION

Adoption of drone and autonomous technology has become the modern arms race and the U.S. has shown itself willing to push to the forefront of these new technologies. This new arms race is unlike the nuclear arms race in that there is no clear first mover, or innovator, advantage. Instead, in the era of Big Data, as shown in the use example of these technologies, the operator that is best

able to use the vast amount of information available to them simultaneously will hold the advantage. The U.S. is making promising steps towards this end and will face new, difficult choices in how to integrate these innovations.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the Associate Instructors for their support and suggestions in exploring this topic.

## REFERENCES

- [1] Greg Allen and Taniel Chan. 2017. *Artificial Intelligence and National Security*. Technical Report, Harvard Kennedy School - Belfer Center for Science and International Affairs, 79 JFK Street, Cambridge, MA 02138.
- [2] Sebastian Anthony. 2013. DARPA shows off 1.8-gigapixel surveillance drone, can spot a terrorist from 20,000 feet. Website. (01 2013). <http://www.extremetech.com/extreme/146909-darpa-shows-off-1-8-gigapixel-surveillance-drone-can-spot-a-terrorist-from-20000-feet>
- [3] Allison Barrie. 2015. 'Star Trek'-style surveillance drone for the US military. Website. (09 2015). <http://www.foxnews.com/tech/2016/09/15/star-trek-style-surveillance-drone-for-us-military.html>
- [4] Richard A. Burgess. 2015. ACTUV Sea Trials Set for Early 2016. Website. (11 2015). <http://science.dodlive.mil/2015/11/09/actuv-sea-trials-set-for-early-2016/>
- [5] Jamie Condliffe. 2017. *A 100-Drone Swarm, Dropped from Jets, Plans Its Own Moves*. resreport. MIT Technology Review.
- [6] Deagel. 2017. RQ-4A Global Hawk. Website. (04 2017). <http://www.deagel.com/Support-Aircraft/RQ-4A-Global-Hawk.a000556001.aspx>
- [7] The Economist. 2017. *Taking Flight*. resreport. The Economist: Technology Quarterly.
- [8] Emily Feng and Charles Clover. 2017. Drone swarms vs conventional arms: Chinafis military debate. Website. (08 2017). <https://www.ft.com/content/302fc14a-66ef-11e7-8526-7b38dcaef614?mhq5j=e7>
- [9] Gartner. 2017. Gartner Says Almost 3 Million Personal and Commercial Drones Will Be Shipped in 2017. Press Release. (02 2017). <https://www.gartner.com/newsroom/id/3602317>
- [10] Dan Gettinger. 2017. *Drones in the Defense Budget*. resreport. Center for the Study of Drones, Bard College.
- [11] Northrop Grumman. 2017. Global Hawk. Website. (2017).
- [12] Derrick Harris. 2015. Google: Our new system for recognizing faces is the best one ever. Website. (03 2015). <http://fortune.com/2015/03/17/google-facenet-artificial-intelligence/>
- [13] Dan Lamothe. 2016. Watch Perdix, the secretive Pentagon program dropping tiny drones from jets. Website. (03 2016). [https://www.washingtonpost.com/news/checkpoint/wp/2016/03/08/watch-perdix-the-secretive-pentagon-program-dropping-tiny-drones-from-jets/?utm\\_term=.0a44c6311045](https://www.washingtonpost.com/news/checkpoint/wp/2016/03/08/watch-perdix-the-secretive-pentagon-program-dropping-tiny-drones-from-jets/?utm_term=.0a44c6311045)
- [14] John Markoff. 2013. Military Lags in Push for Robotic Ground Vehicles. Website. (09 2013). <http://www.nytimes.com/2013/09/24/science/military-lags-in-push-for-robotic-ground-vehicles.html>
- [15] 60 Minutes. 2017. New generation of drones set to revolutionize warfare. Television. (01 2017). Correspondent David Martin.
- [16] Kyle Mizokami. 2017. The Pentagon's Autonomous Swarming Drones Are the Most Unsettling Thing You'll See Today. Website. (01 2017). <http://www.popularmechanics.com/military/aviation/a24675/pentagon-autonomous-swarming-drones/>
- [17] Kris Osborn. 2017. Navy sub-hunting drone ship goes on offense. Website. (01 2017). <https://defensesystems.com/articles/2017/01/11/seahunter.aspx>
- [18] Caroline Reese. 2017. Endeavor Robotics to Provide U.S. Government with Throwaway UGV. Website. (09 2017). <http://www.unmannedsystemstechnology.com/2017/09/endeavor-robotics-provide-u-s-government-throwable-ugv/>
- [19] Tyler Rogoway. 2014. Why The USAF's Massive \$10 Billion Global Hawk UAV Is Worth The Money. Website. (09 2014). <https://foxtrotalpha.jalopnik.com/why-the-usafs-massive-10-billion-global-hawk-uav-was-w-1629932000>
- [20] Adam Stone. 2016. ACTUV on track for Navy success story. Website. (12 2016). <https://www.c4isrn.net/unmanned/uas/2016/12/21/actuv-on-track-for-navy-success-story/>
- [21] Patrick W. Watson. 2017. U.S. Military May Soon Deploy Millions Of Drones, Which Presents A Big Investment Opportunity. Website. (08 2017). <https://www.forbes.com/sites/patrickwwatson/2017/08/02/u-s-military-may-soon-deploy-millions-of-drones-which-presents-a-big-investment-opportunity/#4068439334a8>

```
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
```

```
=====
[2017-11-06 17.38.09] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Missing character: ""
Missing character: ""
Typesetting of "report.tex" completed in 0.9s.
./README.yml
14:81    error    line too long (81 > 80 characters) (line-length)
15:81    error    line too long (83 > 80 characters) (line-length)
16:81    error    line too long (85 > 80 characters) (line-length)
17:81    error    line too long (84 > 80 characters) (line-length)
18:81    error    line too long (95 > 80 characters) (line-length)
19:81    error    line too long (87 > 80 characters) (line-length)
20:81    error    line too long (81 > 80 characters) (line-length)
21:81    error    line too long (81 > 80 characters) (line-length)
22:81    error    line too long (88 > 80 characters) (line-length)
47:81    error    line too long (842 > 80 characters) (line-length)
```

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

```
name: Peter Russell
hid: 334
paper1: Oct 28 17 100%
paper2: 100%
project: in progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
3
wc 334 paper2 3 2816 report.tex
wc 334 paper2 3 2906 report.pdf
wc 334 paper2 3 655 report.bib
```

```
find "
```

---

- 26: Traditionally, drones can be categorized as a re-usable, autonomous vehicle either in the air or on the ground. In the air, these are known as "Unmanned Aerial Vehicles", or UAVs, and on the ground, "Unmanned Ground Vehicles", or UGVs. More recently, this has expanded to USVs for "Unmanned Surface Vehicles" and UUVs for "Unmanned Underwater Vehicles." Our focus will remain primarily on the former two types, UAVs and UGVs since these have been the most long-standing types.
- 59: Eventually, these swarm drones are expected to have the capability to be aggregated together by the thousands and carry out overwhelming and confusing attacks on enemies. It has been properly described as the "difference between a wolf pack and just little wolves\cite{ftswarm}."

```
passed: False
```

```
find footnote
```

---

```
passed: True
```

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

---

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

passed: True

floats

---

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

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

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

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

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

Label/ref check

passed: True

When using figures use columnwidth

[width=1.0\columnwidth]

do not change the number to a smaller fraction

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

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

```
bibtex_empty_fields
```

---

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

```
find ""
```

---

```
passed: True
```

---

```
ascii
```

---

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

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

```
find newline
```

---

```
passed: True
```

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

```
find cite {
```

---

```
passed: True
```

# Big Health Data from Wearable Electronic Sensors (WES) and the Treatment of Opioid Addiction

Sean M. Shiverick

Indiana University Bloomington

smshiver@indiana.edu

## ABSTRACT

Wearable electronic sensors (WES) and mobile health applications can be used to collect vital health data to supplement traditional forms of treatment for opioid addiction and may be used to predict risk factors related to overdose death.

## KEYWORDS

Health Informatics, Wearable Sensors, Addiction Treatment, i535, HID335

## 1 INTRODUCTION

In the increasingly connected digital age, personal electronic devices are generating huge volumes of data with important applications for health informatics. Wearable electronic sensors (i.e., *wearables*) and fitness monitors (e.g., FitBit, iWatch) can record our movements and vital physiological measures such as heart rate, temperature, and blood pressure [7]. Consumers are using wearables to self-monitor stress and hypertension, and wearable sensors can be used to help track recovery following medical procedures such as surgery [2]. The development of personalized health care models are also enabling individuals to self-monitor and manage their own health in partnership with care providers. This paper explores approaches to using personal electronic devices and wearable sensors for the treatment of addiction disorders and the prevention of drug overdose. Past research has shown that *Mobile Health* platforms have been used to address prescription medication abuse in several ways: (a) monitor patient health conditions at any time and remotely, (b) monitor medication consumption, and (c) connect patients with health care providers and treatment services [19]. The following review of the literature shows that wireless digital technologies and smartphone applications are effective at providing health data in real time and can assist patients in recovery to resist physical cravings, prevent relapse, and access treatment support. Mobile applications can play an important role in addressing the opioid epidemic by supplementing traditional approaches to addiction treatment and recovery.

### 1.1 The Opioid Epidemic: Medication Abuse and Addiction

The abuse of prescription opioid medication in the U.S. has become a major health crisis that the Department of Health and Human Services (HHS) has described as an epidemic [20]. Approximately 2 million Americans were dependent on or abused prescription opioids (e.g., oxycodone, hydrocodone) in 2014 [9]. Overdose deaths from prescription opioids has quadrupled since 1999, resulting in more than 180,000 deaths between 1999 to 2015. Figure 1 shows that the dramatic increase in overdose deaths in the U.S. between

2000 and 2016 are from synthetic opioids (other than methadone), natural and semi-synthetic opioids, and heroin [14]. Of the estimated 64,000 drug overdose deaths in 2015, over 20,000 were from fentanyl and other synthetic opioid analogs. Public health agencies are implementing comprehensive efforts to address four major risk areas of prescription opioid abuse, overdoses, and deaths: (i) Increasing knowledge of opioid abuse and improving decisions among medication prescribers, (ii) Reducing inappropriate access to opioids, (iii) Increasing effective overdose treatment, (iv) Providing substance-abuse treatment to persons addicted to opioids. The opioid epidemic is complex, with multiple and interacting causal factors. To understand how technological interventions can play a role in mitigating the crisis, it is necessary to consider the nature of addiction itself and various approaches to treatment.

[Figure 1 about here.]

**1.1.1 Drug Addiction and Treatment.** For millions of people struggling with substance abuse and dependency in the U.S., addiction and relapse are chronic health conditions [4]. Drug addiction has many similar characteristics to other chronic medical illnesses; however, there are unique challenges to the treatment of addiction illnesses. For example, drug addicted patients undergo intense detoxification in rehabilitation treatment programs, which reduces their drug tolerance, and then are released back into the same environment associated with their drug use, putting them at greater risk for relapse and potential drug overdose. The lack of continuity in the treatment of addiction disorders leaves persons in recovery at high risk of relapse for substance use and abuse. Second, individuals with severe addiction disorders end up at emergency rooms for care following acute intoxication, often following law enforcement interventions. Emergency personnel are very competent at crisis interventions for drug overdose, but lack resources to evaluate severe addiction disorders or provide follow-up. Furthermore, addicted individuals seeking treatment often relapse at night or on weekends when treatment centers are not open. Various theories of addiction and relapse have been proposed. According to the classical conditioning model, situational cues or events can elicit a motivational state underlying relapse to drug use. A slightly more complex model suggests that addictive behavior can be reinstated after extinction of dependency by exposure to drugs, drug-related cues, or environmental stressors [15]. Understanding that a user's affective (i.e., motivational) response to cues in the environment can lead to relapse and drug use are key to developing strategies for prevention and treatment.

## 1.2 Technology-Based Interventions for Addiction Treatment

Technology-based interventions have been used for drug addiction assessment, treatment, prevention, and recovery [12]. In terms of assessment, data about substance use can be obtained from mobile cell phone reporting outside of treatment settings. Web-based approaches to treatment have been implemented online to improve behavioral and psychosocial functioning for addicted individuals in recovery [13]. For example, the *Therapeutic Education System* (TES) is a self-directed, web-based interactive treatment program consisted of 65 training modules that focused on cognitive-behavioral skills and psychosocial functioning (family/social relations). This online approach helped to increase access to treatment for individuals in rural areas, and included an optional contingency management module. A computer based *Training in Cognitive Behavioral Therapy* (CBT) program was found to enhance treatment outcomes when provided in conjunction with traditional substance abuse treatment, and helped improve coping skills and decision-making skills [6]. In evaluating the effectiveness of mobile applications for addiction treatment, several questions remain to be answered: First, if mobile applications are regarded primarily as supplements to traditional therapeutic treatment, can their effectiveness be evaluated independently from the approach used in treatment? Second, over what time period can the benefits of mobile applications be observed? Research evidence suggests that the benefits of mobile interventions may be limited to 12 or 15 weeks [16]. It is unclear whether individuals struggling from addiction would continue to use mobile treatment applications in the long term, beyond a limited course of treatment.

**1.2.1 Mobile-Based Applications.** Mobile applications have been used for monitoring and treatment of substance abuse and addiction disorders for several decades [4]. Early applications included the use of electronic pagers (i.e., beepers) for experience sampling with paper-based assessments that generated data about daily life behavior and experiences [16]. In the 1990s, programmable personal digital assistants (e.g., palm-pilot) enabled collection of data electronically, and subsequent mobile research tools facilitated the collection of information about psychological factors (e.g., daily stressors, emotional states, thoughts) and other variables related to addiction (e.g., craving, contextual cues, actual substance use). Assessments performed several times throughout the day (commonly, every 2 to 4 hours) allowed for analysis of the daily fluctuations of these symptoms and features. Historically, addiction research has faced some unique challenges that the use of mobile technologies may help to overcome. Methodological aspects of traditional research using retrospective, cross-sectional, or longitudinal assessments (over periods of weeks, months, or years) have been problematic for investigating risk factors including behaviors and symptoms (severe physiological cravings, withdrawal, and substance use) that can span a relatively short time. An additional factor is the co-morbidity, or co-occurrence, of substance use disorders (SUDs) with other psychological disorders, such as anxiety and mood disorders. For example, the *self-medication* model has commonly been used to explain the association between alcohol abuse as an effort by an individual to reduce or cope with a high degree of anxiety (or depression). It has also been challenging for

researchers to capture the role of environmental or contextual cues (e.g., people, places, things) associated with substance abuse, which can act as triggers of relapse for individuals in recovery.

**Smartphone Applications.** Continued care is an important ingredient for recovery from addiction that involves monitoring, outreach, planning, case management, and social support [11]. Smartphone applications can help individuals in recovery to monitor cravings at critical points in daily life, track contextual cues associated with substance use, and provide outreach to support services. A team of researchers at the University of Wisconsin evaluated the effectiveness of a smartphone application called *Addiction Comprehensive Health Enhancement Support System* (A-CHESS), designed to provide recovery support for patients leaving a residential alcohol treatment center [10]. A-CHESS provided anytime, anywhere access to support services in audio-visual format, GPS monitoring and warnings for risky locations related to past substance use, and communication with counselors. Over an 8-month period (and 4 month follow-up), patients who used the A-CHESS intervention reported fewer risky drinking days, on average, per month than patients in a comparable control group. The findings provide evidence that the smartphone intervention was effective at treating a critical behavioral measure for treatment of alcohol use disorder (AUD). The methods described in this study could be extended by re-purposing built-in smartphone sensors to record physiological measures related to opioid usage, and communicate data to health care providers or treatment specialists to initiate interventions for opioid addiction [11].

[Figure 2 about here.]

## 1.3 Medication Adherence and Abuse Monitoring System

Mobile health applications can be used to monitor medication adherence and as an advanced warning system for potential abuse of prescription medication [18]. Medication abuse can consist of higher medication dosages or rapid escalation of a prescribed dosage, and the general goal of a prediction model is to analyze patient data for sudden changes in medication consumption. Figure 2 illustrates several steps in a process and decision support structure for a medication monitoring system, with adjustable parameters, such as the threshold for abuse (e.g., greater than N doses in X hours) [19]. A major challenge for measuring medication abuse is obtaining reliable information from potentially addicted individuals based on self report data. Ideally, information on medication consumption and adherence can be obtained from multiple sources. Addiction is a complex behavior that involves a variety of factors, including: demographics (e.g., age, gender), past history, comorbidity with other disorders, family support, social influence, employment status, and patient motivation. Figure 3 shows a model architecture of a system for monitoring potential abuse where dose information is provided via a smartphone application, relayed via wireless cellular network to analytic models that measure changes in medication consumption, relays reports to support treatment services for possible interventions, and to a smart medication box that dispenses medication. In order to function successfully a medication abuse monitoring system depends on the collection of reliable information, including data from wearable sensors that can directly

measure physiological changes (e.g., heartrate, blood pressure, respiration, temperature) related to changes in medication usage. In the context of prescription opioid abuse, a medication monitoring system could be very beneficial in anticipating opioid dependency and preventing accidental death from medication overdose.

[Figure 3 about here.]

#### 1.4 Mobile Detection with Wearable Biosensors

Portable biosensors can provide a continuous stream of data on the timing, location, context, and duration of drug use by individuals in treatment. In a small pilot study, researchers used an Affectiva Q sensor to measure electrodermal activity (EDA), skin temperature, and acceleration (8 recordings per second), in a sample of N = 4 patients during the administration of opioid medication in an emergency room setting [5]. Table 1 provides a summary of the participant characteristics. The biosensor was worn on the wrist and was similar in size and dimensions to a wristwatch or fitbit health monitor. The results showed an increase in EDA associated with intravenous opioid injection that was detected by the biosensors. In addition, there was some indication that the physiological response to opioids varied according to individual drug tolerance; patients with higher opioid tolerance showed less EDA response than patients with low tolerance. The findings provide evidence to support the use of wearable sensors to detect drug use in real time, in a controlled environment. An important limitation of the study is the small sample size, which reduces the generalizability of the findings to a broader population. The authors also acknowledged that psychological or physiological stress can produce alterations in EDA, skin temperature, and acceleration, and therefore this could not be ruled out as an alternative explanation for the findings. The results are promising, however, and encourage efforts to explore the effectiveness of wearable biosensors in the context of environments associated with substance use.

[Table 1 about here.]

#### 1.5 Emerging Sensor Technologies

Wearable wireless sensors have been used to study physiological responses, activity, and social behavior in non-human primates in the form of a fitted vest and using a mobile phone with blue tooth protocol to collect data in real time. Figure 4 shows sample ambulatory data from a rhesus macaque recorded from a wearable wireless sensor for 11 hours inside a large group primate cage [8]. Data was recorded on a custom Android software application, which captured measures of EDA, heart rate (HR), temperature, and acceleration. The goals of this study were to measure associations between physiological measures and social behavior in primates; however, this practical application of sensor technology demonstrated a system that was relatively low-cost, highly portable, scalable, and simple to use. Future research could explore the development of a similar system modified for use with humans to collect data on physiological measures from addicted individuals in naturalistic settings.

[Figure 4 about here.]

**1.5.1 LoRa Backscatter: Enabling Ubiquitous Connectivity.** Emerging technologies, such as long range (LoRa) backscatter, have the potential to extend the boundaries of wireless connectivity. Existing

radio technologies (e.g., WIFI, ZigBee, SigFox, LTE-M) provide reliable long range coverage, but consume energy and would be costly to expand to large scale implementation; however, LoRa backscatter is a smaller, low-cost, low-power alternative with extended range between an RF source and receiver of approximately 475 meters (i.e., yards) [17]. Table 1 shows the sensitivity and supported data rates for different communication technologies and feasibility of different power sources. LoRa backscatter performed best overall, in terms of sensitivity (-149 dBm), supporting bit rates of 18 pbs to 37.5 kbps, providing whole home coverage, and capable of being powered by button cell, tiny solar cell, or printed battery. LoRa backscatter uses chirp spread activation (CSS) that can synthesize continuous frequency modulated chirps; a limitation is that backscatter is drowned out by noise and the RF source. The LoRa backscatter system was tested in various deployments: across three floors of a 4800 square- foot house, a single floor of 13,000 square foot office building, and on a one-acre farm. Figure 5 shows the layout of the house with the RF source (TX) on the second floor and receiver in the basement (RX); the plot shows the system achieved RSSI values greater than -144 dBm, with reliable wireless coverage throughout the house, and rates sufficient for temperature sensors that transmit small packages. The system was also implemented in the form flexible epidermal patch sensor shown in Figure 6, that provided reliable connectivity across a 3,300 square foot atrium with RSSI greater than -132 dBm. LoRa backscatter provides a compact, energy-efficient, and affordable wireless transmission system that can be extended to scale at reasonable cost. This system could possibly transmission of biometric data from wearable sensors to capture health information from addicted individuals in treatment.

[Table 2 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

**1.5.2 Graphene Electronic Tattoo sensors.** Wearable, tattoo-like epidermal sensors allow for continuous, ambulatory monitoring of biometric signals from the heart, muscles, and brain, outside of hospitals and clinical lab settings [3]. A team of researchers at the University of Texas at Austin designed the graphene electronic tattoo (GET) as a long term wearable sensor that can be directly laminated on human skin, and can remain functional for several days with a liquid bandage cover [1]. Graphene is the thinnest electrically conductive material that is biocompatible, stable, and mechanically robust. The “GET is fabricated through a simple ‘wet transfer, dry patterning’ process directly on tattoo paper, allowing it to be transferred on human skin exactly like a temporary tattoo, except the sensor is transparent”(p.8)[1]. As depicted in Figure 7, the GET sensor is is flexible, stretchable, and transparent, and less than a sub-micrometer in thickness (463 +/- 30 nm). GET has been used successfully to measure electrocardiograph (EKG), electromyogram (EMG), electrocephalograph (EEG) signals, as well as skin temperature and skin hydration. After use, the GET can be easily removed by peeling it from the skin. A future step in the development of GET is to include an antenna to the design so that signals can be beamed off the device to a smartphone application or computer. The thin, flexible, resilient tattoo biosensor provides a durable, unobtrusive tool for collecting physiological data, and

could be used to detect physical changes due to drug withdrawal in addicted individuals.

[Figure 7 about here.]

## 2 CONCLUSION

*Can Technological Applications Reduce Opioid Addiction?* The abuse of prescription medication in the U.S. has led to opioid addiction at levels of epidemic proportion. Technological interventions can play a role in addressing this crisis as a supplement to conventional forms of addiction treatment. Mobile health applications can help monitor potential medication abuse and connect individuals with treatment services. An important limitation of data based addiction interventions is the difficulty of obtaining reliable information about medication consumption based on self-reports from potentially addicted individuals. The literature reviewed indicates that wearable sensors are an effective way to measure vital health data in real time and remotely. Providing individuals in recovery with vital health data may help them to resist physical cravings and prevent relapse. Another limitation of treatment approaches is that, after detoxification, individuals in recovery are released back into the environmental settings associated with their drug use, putting them at risk for potential relapse and possible overdose. Recent advances in signal technologies such as LoRa Backscatter and Graphene tattoo sensors can lead to the more efficient collection of biometric information and cost effective transmission of health data for subsequent analysis. The opioid addiction epidemic is a complex phenomenon, with both physical and sociological contributing factors. Technological interventions will increase the amount of data about addicted individuals and relevant risk factors that may be used to predict opioid overdose death; however, it will not address the environmental factors that lead to addiction. Despite increased awareness of the potential for prescription medication abuse, Table 1 shows the rate of overdose deaths is growing more rapidly for heroin and synthetic opioids such as fentanyl compared to conventional prescription opioid medication. The implication of this is that individuals who may become addicted to prescribed medication may go on to abuse illicit or synthetic opioids, in non-clinical, unsupervised, and unregulated settings. Big data offers potential for transforming health care and addition treatment. Increasing levels of data about opioid addiction ultimately may not be sufficient to prevent or decrease rates of overdose death if the availability of illicit and synthetic opioids remains high.

## ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, the Assistant Instructors, Juliette Zurick and others, and anonymous reviewers who helped to improve this report.

## REFERENCES

- [1] Shideh Kabiri Ameri, Rebecca Ho, Hongwoo Jang, Li Tao, Youhua Wang, Liu Wang, David M. Schnyer, Deji Akinwande, and Nanshu Lu. 2017. Graphene Electronic Tattoo Sensors. *ACS Nano* 11, 8 (2017), 7634–7641. <https://doi.org/10.1021/acsnano.7b02182> arXiv:<http://dx.doi.org/10.1021/acsnano.7b02182> PMID: 28719739.
- [2] Louis Atallah, Gareth G. Jones, Raza Ali, Julian J. H. Leong, Benny Lo, and Guang-Zhong Yang. 2011. Observing Recovery from Knee-Replacement Surgery by Using Wearable Sensors. In *Proceedings of the 2011 International Conference on Body Sensor Networks (BSN '11)*. IEEE Computer Society, Washington, DC, USA, 29–34. <https://doi.org/10.1109/BSN.2011.10>
- [3] Katherine Bourzac. 2017. Graphene Temporary Tattoo Tracks Vital Signs. online. (Jan. 2017). <https://spectrum.ieee.org/nanoclast/semiconductors/nanotechnology/graphene-temporary-tattoo> IEEE Spectrum.
- [4] E.W. Boyer, D. Smelson, R. Fletcher, D. Ziedonis, and Picard R. W. 2010. Wireless Technologies, Ubiquitous Computing and Mobile Health: Application to Drug Abuse Treatment and Compliance with HIV Therapies. *Journal of Medical Toxicology* 6, 2 (2010), 212–216. <https://doi.org/10.1007/s13181-010-0080-z>
- [5] Stephanie Carreiro, David Smelson, Megan Ranney, Keith J. Horvath, R. W. Picard, Edwin D. Boudreaux, Rashelle Hayes, and Edward W. Boyer. 2015. Real-Time Mobile Detection of Drug Use with Wearable Biosensors: A Pilot Study. *Journal of Medical Toxicology* 11, 1 (Oct. 2015), 73–79. <https://doi.org/10.1007/s13181-014-0439-7>.
- [6] K.M. Carroll, S.A. Ball, S. Martino, and et al. 2008. Computer-assisted delivery of cognitive-behavioral therapy for addiction: a randomized trial of CBT4CBT. *Am J Psychiatry* 165, 7 (2008), 881f1?8. <https://doi.org/10.1176/appi.ajp.2008.07111835>
- [7] Melinda Gomez Michael Schwartz David Metcalf, Sharlin TJ. Milliard. 2016. Wearables and the Internet of Things for Health. *IEEE Pulse* (Oct. 2016). <https://pulse.embs.org/september-2016/wearables-internet-of-things-iot-health/>
- [8] Richard Ribon Fletcher, Ken ichi Amemori, Matthew Goodwin, and Ann M. Graybiel. 2012. Wearable wireless sensor platform for studying autonomic activity and social behavior in non-human primates. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, Annual International Conference of the IEEE (Ed.). IEEE, IEEE, San Diego, CA, USA. <https://doi.org/10.1109/EMBC.2012.6346855>
- [9] Centers for Disease Control and Prevention. 2017. Prescription Opioid Overdose Data. online. (Oct. 2017). <https://www.cdc.gov/drugoverdose/data/overdose.html>
- [10] D.H. Gustafson, F.M. McTavish, M.-Y. Chih, A.K. Atwood, R.G. Johnson, M. Boyle, and M. ... Shah. 2014. A smartphone application to support recovery from alcoholism: A randomized controlled trial. *JAMA psychiatry* 71, 5 (May 2014), 566–572. <https://doi.org/10.1001/jamapsychiatry.2013.4642>
- [11] 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>
- [12] Lisa A. Marsch. 2012. Leveraging technology to enhance addiction treatment and recovery. *Journal of Addictive Diseases* 31, 3 (2012), 313–318. <https://doi.org/10.1080/10550887.2012.694606>
- [13] L. A. Marsch and J. Dallery. 2012. Advances in the Psychosocial Treatment of Addiction: The Role of Technology in the Delivery of Evidence-Based Psychosocial Treatment. *The Psychiatric Clinics of North America* ;35(2): doi: 2 (2012), 481–493. <https://doi.org/10.1016/j.psc.2012.03.009>
- [14] 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>
- [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] J. Swendsen. 2016. Contributions of mobile technologies to addiction research. *Dialogues Clinical Neuroscience* (2016). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969708/>
- [17] Vamsi Talla, Mehrdad Hessar, Bryce Kellogg, Ali Najafi, Joshua R. Smith, and Shyamnath Gollakota. 2017. LoRa Backscatter: Enabling The Vision of Ubiquitous Connectivity. *CoRR abs/1705.05953* (2017). <https://arxiv.org/abs/1705.05953>
- [18] 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>
- [19] Upkar Varshney. 2014. Mobile Health: Medication Abuse and Addiction. In *Proceedings of the 4th ACM MobiHoc Workshop on Pervasive Wireless Healthcare (MobileHealth '14)*. ACM, New York, NY, USA, 37–42. <https://doi.org/10.1145/2633651.2633656>
- [20] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. 2014. Medication-Assisted Therapies: Tackling the Opioid-Overdose Epidemic. *New England Journal of Medicine* 370, 22 (2014), 2063–2066. <https://doi.org/10.1056/NEJMmp1402780> PMID: 24758595.

## 3 ISSUES

DONE:

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

### 3.1 Assignment Submission Issues

DONE:

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

## 3.2 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

## 3.3 Formatting

DONE:

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

DONE:

Other formatting issues

## 3.4 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}

## 3.5 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

## 3.6 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

## 3.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

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

DONE:

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

DONE:

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

DONE:

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

### 3.8 Details about the Figures and Tables

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

DONE:

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

e.g.

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

#### LIST OF FIGURES

1	Drugs Involved in U.S. Overdose Deaths from 2000 to 2016, National Institute on Drug Addiction (NIDS) [14]	8
2	Process and Decision Support for Abuse Monitoring System [19]	9
3	Architecture for Abuse Monitoring System [19]	10
4	Sample Ambulatory Data from Rhesus Macaque Recorded on Wearable Sensor for 11+ hours Inside Large Primate Cage Facility [8]	11
5	Home Deployment of LoRa backscatter pakcets across 4,800 sq. ft. House Apread Across Three Floors [17]	12
6	LoRa Backscatter Epidermal Patch [17]	13
7	Graphene Electronic Tatoo Biosensor [1]	13

## Drugs Involved in U.S. Overdose Deaths, 2000 to 2016

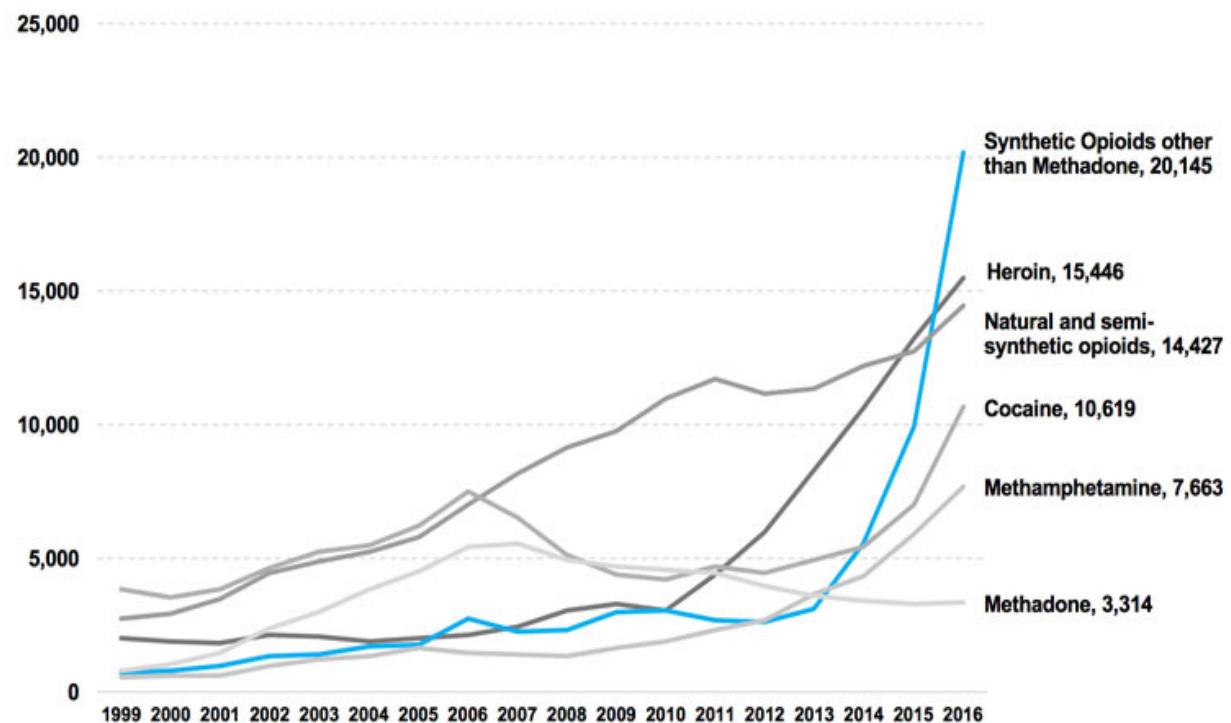
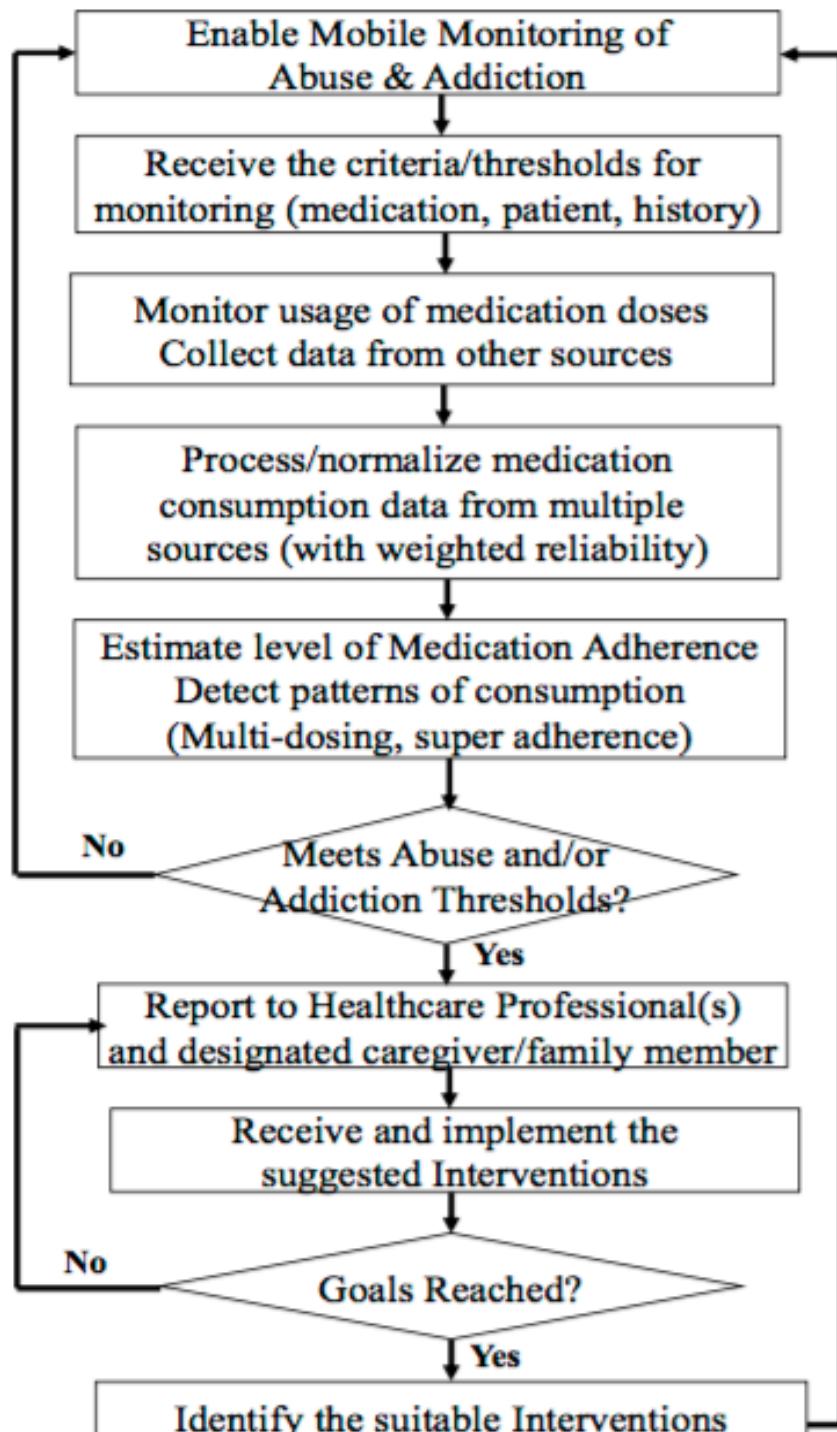
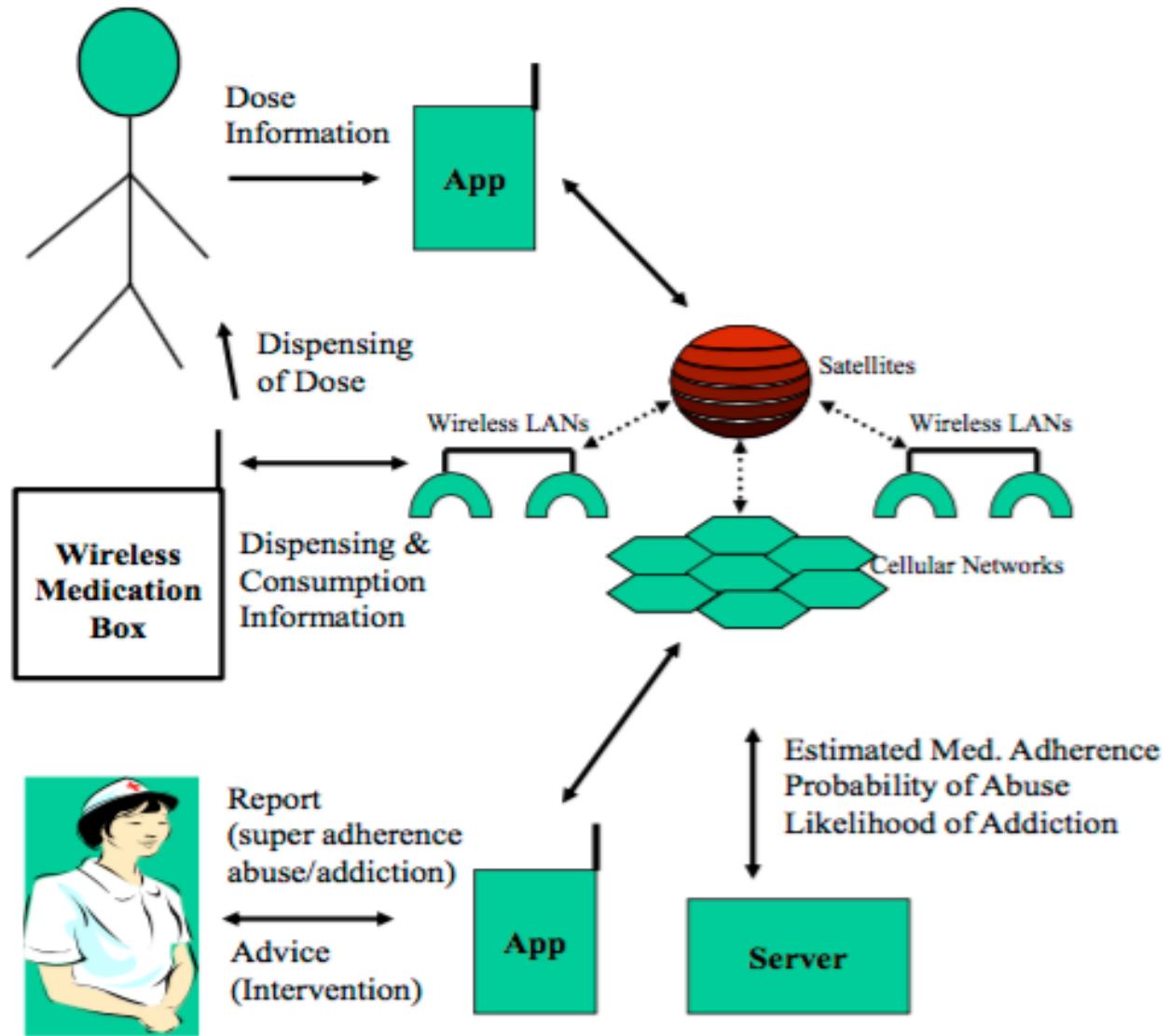


Figure 1: Drugs Involved in U.S. Overdose Deaths from 2000 to 2016, National Institute on Drug Addiction (NIDS) [14]



(b) Process and Decision Support



(a) Architecture of the Abuse Monitoring System

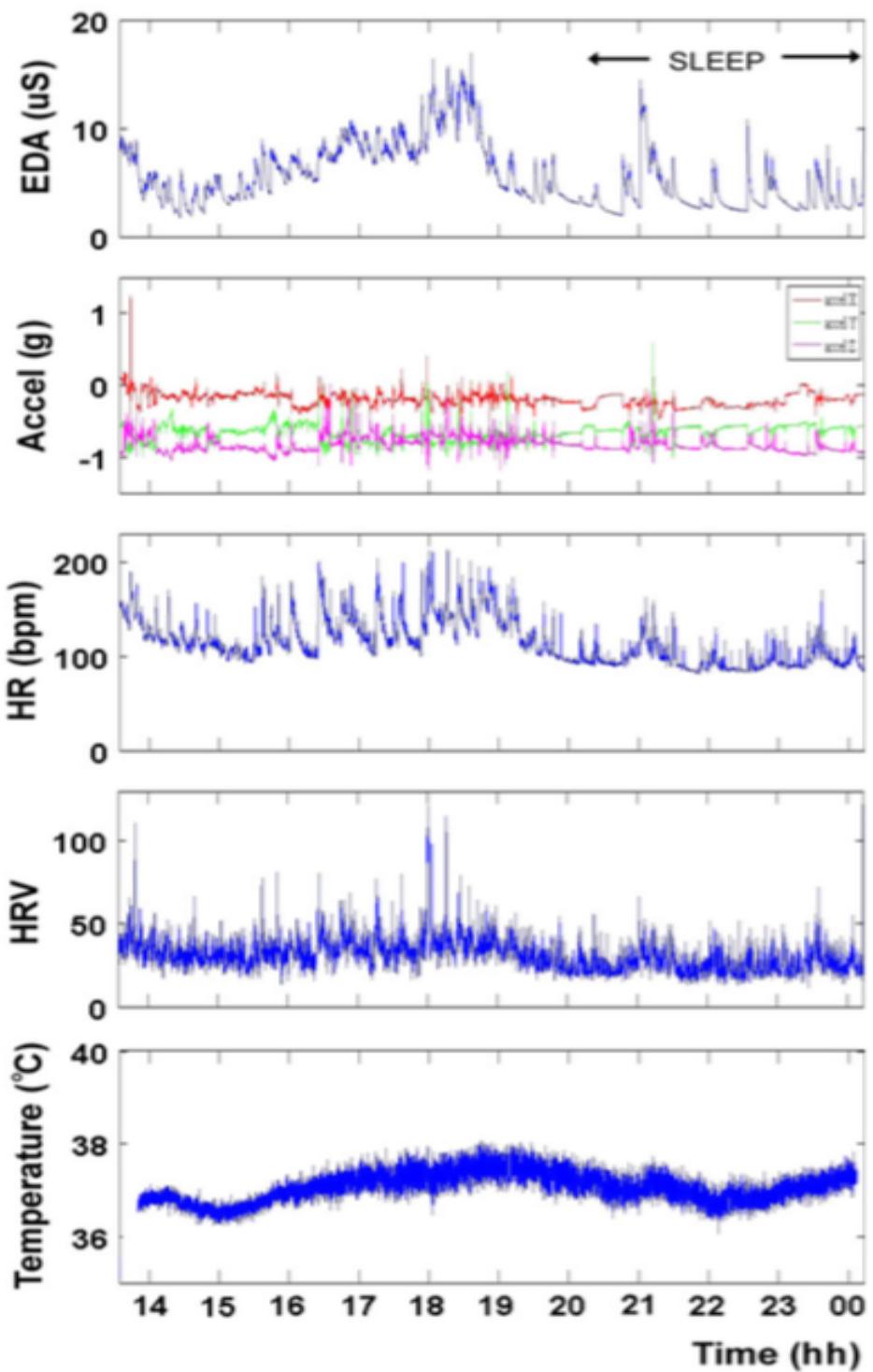


Figure 4: Sample Ambulatory Data from Rhesus Macaque Recorded on Wearable Sensor for 11+ hours Inside Large Primate Cage Facility [8]  
286

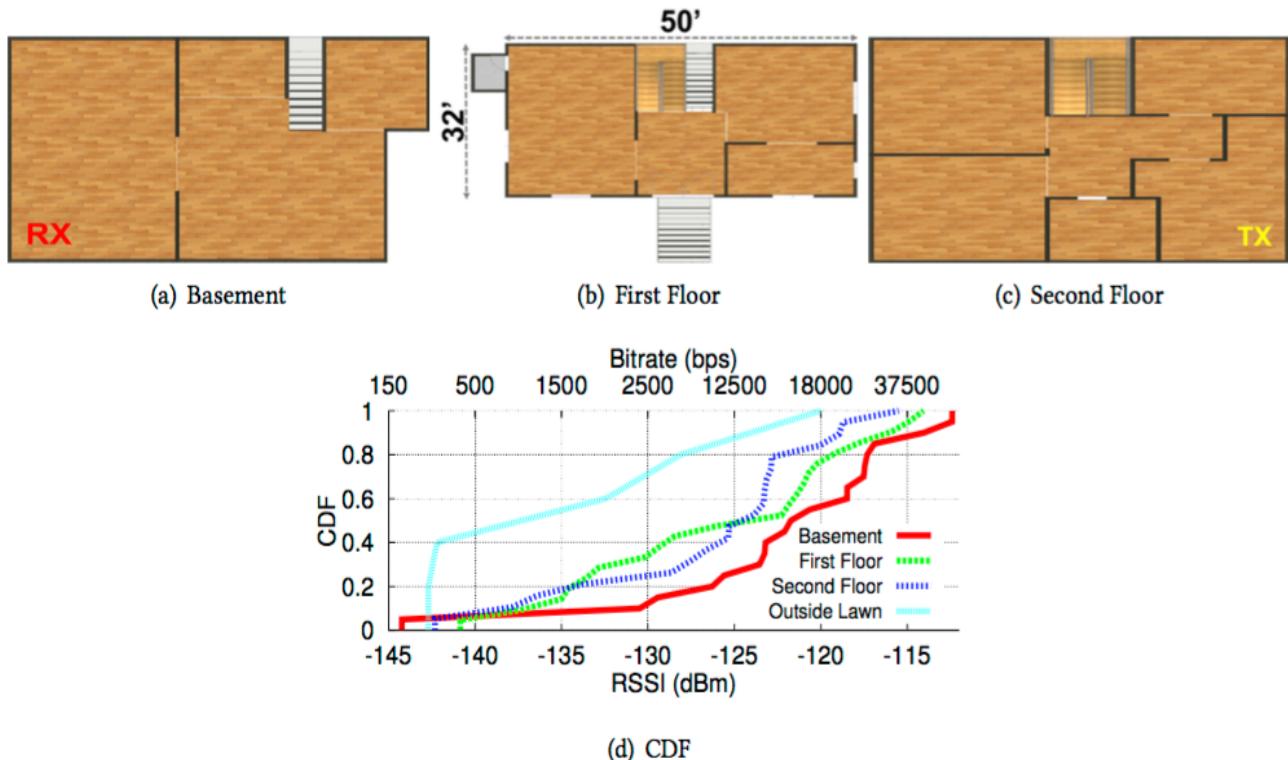


Figure 5: Home Deployment of LoRa backscatter packets across 4,800 sq. ft. House spread Across Three Floors [17]



Figure 6: LoRa Backscatter Epidermal Patch [17]



Figure 7: Graphene Electronic Tattoo Biosensor [1]

LIST OF TABLES

1	Summary of Participant Characteristics in Pilot study [5]	15
2	Comparison of Wireless Communication Technologies [17]	15

**Table 1: Summary of Participant Characteristics in Pilot study [5]**

Patient	Age	Gender	History of Use	Intervention	Pre-EDA	Post-EDA
1	82	Male	Opioid naive	4 mg morphine	4.5	60.0
2	47	Male	Recent short-term	1 mg hydromorphone	3.4	12.2
3	43	Female	Chronic opioid use	1 mg hydromorphone	0.2	0.2
4	72	Male	Chronic opioid use	4 mg morphine	0.9	1.6

**Table 2: Comparison of Wireless Communication Technologies [17]**

Technology	Sensitivity	Data Rate	Home Coverage	Button Cell	Tiny Solar Cell	Printed Battery
Wi-Fi (802.11 b/g)	-95 dBm	1-54 Mbps	yes	no	no	no
LoRa	-149 dBm	18 bps-37.5 kbps	yes	no	no	no
Bluetooth	-97 dBm	1-2 Mbps	no	no	no	no
SigFox	-126 dBm	100 bps	yes	no	no	no
Zigbee	-100 dBm	250 kbps	yes	no	no	no
Passive Wi-Fi	-95 dBm	1-11 Mbps	no	yes	yes	yes
RFID	-85 dBm	40-640 kbps	no	yes	yes	yes
LoRA Backscatter	-149 dBm	18 bps-37.5 kbps	yes	yes	yes	yes

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 "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3085 of file ACM-Reference-Format.bst  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3085 of file ACM-Reference-Format.bst  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3131 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3131 of file ACM-Reference-Format.bst  
Warning--numpages field, but no articleno or eid field, in atallah11  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3229 of file ACM-Reference-Format.bst  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--unrecognized DOI value [doi:10.1007/s13181-010-0080-z]  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--no number and no volume in metcalf16  
Warning--page numbers missing in both pages and numpages fields in metcalf16  
Warning--page numbers missing in both pages and numpages fields in fletcher12  
Warning--no number and no volume in johnson11  
Warning--page numbers missing in both pages and numpages fields in johnson11  
Warning--no number and no volume in swedenson16  
Warning--page numbers missing in both pages and numpages fields in swedenson16  
Warning--page numbers missing in both pages and numpages fields in talla17  
Warning--numpages field, but no articleno or eid field, in Varshney14  
(There were 12 error messages)  
make[2]: \*\*\* [bibtex] Error 2

latex report

```
=====
```

```
[2017-11-06 17.38.15] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflatex)
Label `tab:freq' multiply defined.
Missing character: ""
bookmark level for unknown defaults to 0.
The anchor of a bookmark and its parent's must not be the same. Added a new anchor.
Float too large for page by 51.24545pt.
Float too large for page by 51.24545pt.
Float too large for page by 62.24545pt.
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 1.8s.
./README.yml
13:5      error    wrong indentation: expected 8 but found 4 (indentation)
15:5      error    wrong indentation: expected 8 but found 4 (indentation)
39:5      error    wrong indentation: expected 8 but found 4 (indentation)
41:5      error    wrong indentation: expected 8 but found 4 (indentation)
46:80     error    trailing spaces (trailing-spaces)
47:77     error    trailing spaces (trailing-spaces)
48:79     error    trailing spaces (trailing-spaces)
49:35     error    trailing spaces (trailing-spaces)
56:4      error    wrong indentation: expected 4 but found 3 (indentation)
56:17     error    trailing spaces (trailing-spaces)
58:4      error    wrong indentation: expected 7 but found 3 (indentation)
60:4      error    wrong indentation: expected 7 but found 3 (indentation)
77:70     error    trailing spaces (trailing-spaces)
81:33     error    trailing spaces (trailing-spaces)
83:81     error    line too long (91 > 80 characters) (line-length)
```

```
=====
```

Compliance Report

```
=====
```

name: Sean Shiverick  
hid: 335

```
paper1: 10/25/17 100%
paper2: 100%
project: in progress
```

```
yamlcheck
```

---

```
wordcount
```

---

```
15
wc 335 paper2 15 3431 report.tex
wc 335 paper2 15 4689 report.pdf
wc 335 paper2 15 1096 report.bib
```

```
find "
```

---

```
passed: True
```

```
find footnote
```

---

```
passed: True
```

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

---

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

```
passed: True
```

```
floats
```

---

```
78: \begin{figure}![ht]
79: \centering\includegraphics[width=\columnwidth]{images/Figure1.pdf}
82: }\label{f:Figure1}
196: \begin{figure}![ht]
197: \centering\includegraphics[width=\columnwidth]{images/Figure2.pdf}
    }
200: }\label{f:Figure2}
233: \begin{figure}![ht]
234: \centering\includegraphics[width=\columnwidth]{images/Figure3.pdf}
    }
```

```

236: }\label{f:Figure3}
265: \begin{table}
267: \label{tab:freq}
298: \begin{figure} [!ht]
299: \centering\includegraphics[width=\columnwidth]{images/Figure4.pdf}
    }
303: \label{f:Figure4}
337: \begin{table}
339: \label{tab:freq}
356: \begin{figure} [!ht]
357: \centering\includegraphics[width=\columnwidth]{images/Figure5.pdf}
    }
360: \label{f:Figure5}
363: \begin{figure} [!ht]
364: \centering\includegraphics[width=\columnwidth]{images/Figure6.pdf}
    }
366: \label{f:Figure6}
395: \begin{figure} [!ht]
396: \centering\includegraphics[width=\columnwidth]{images/Figure7.pdf}
    }
398: \label{f:Figure7}

```

figures 7  
tables 2  
\includegraphics 7  
labels 9  
refs 0  
floats 9

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

Label/ref check  
62: more than 180,000 deaths between 1999 to 2015. Figure 1 shows that  
the dramatic  
210: Figure 2 illustrates several steps in a process and decision  
support structure  
219: status, and patient motivation. Figure 3 shows a model  
architecture of a system  
246: medication in an emergency room setting \cite{carreiro15}. Table  
1 provides a  
285: in real time. Figure 4 shows sample ambulatory data from a rhesus  
macaque  
314: (i.e., yards) \cite{talla17}. Table 1 shows the sensitivity and

```
    supported data
324: one-acre farm. Figure 5 shows the layout of the house with the RF
      source (TX)
329: epidermal patch sensor shown in Figure 6, that provided reliable
      connectivity
382: As depicted in Figure 7, the GET sensor is is flexible,
      stretchable, and
427: for prescription medication abuse, Table 1 shows the rate of
      overdose deaths
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
```

```
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael J. Ziedonis, D. and Picard R. W. Fletcher, R. and Smelson, D. and Boyer, E.W." while executing---line 3085 of file ACM-Reference-Format.bst
```

```
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael J. Ziedonis, D. and Picard R. W. Fletcher, R. and Smelson, D. and Boyer, E.W." while executing---line 3085 of file ACM-Reference-Format.bst
```

```
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael J. Ziedonis, D. and Picard R. W. Fletcher, R. and Smelson, D. and Boyer, E.W." while executing---line 3085 of file ACM-Reference-Format.bst
```

```
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D., and Picard R. W. Fletcher, R. and Smelson, D. and Boyer, E.W." while executing---line 3085 of file ACM-Reference-Format.bst
```

Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3085 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3131 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3131 of file ACM-Reference-Format.bst  
Warning--numpages field, but no articleno or eid field, in atallah11  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3229 of file ACM-Reference-Format.bst  
Name 4 in "Boyer, E.W. and Smelson, D. and Fletcher, R. and Ziedonis, D, and Picard R. W while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--unrecognized DOI value [doi:10.1007/s13181-010-0080-z]  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Too many commas in name 1 of "David Metcalf, Sharlin T.J. Milliard, Melinda Gomez, Michael while executing---line 3229 of file ACM-Reference-Format.bst  
Warning--no number and no volume in metcalf16  
Warning--page numbers missing in both pages and numpages fields in metcalf16  
Warning--page numbers missing in both pages and numpages fields in fletcher12  
Warning--no number and no volume in johnson11  
Warning--page numbers missing in both pages and numpages fields in johnson11  
Warning--no number and no volume in swedenson16  
Warning--page numbers missing in both pages and numpages fields in swedenson16  
Warning--page numbers missing in both pages and numpages fields in talla17  
Warning--numpages field, but no articleno or eid field, in Varshney14  
(There were 12 error messages)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

The following tests are optional

=====

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

find newline

-----

passed: True

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

find cite {

-----

passed: True

# Natural Language Processing (NLP) to Analyze Human Speech Data

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

Extracting meaningful information from large volumes of unstructured human language and deriving sense out of this information is a challenging *Big Data* application. Processing natural language and converting it into meaningful information is a complex task. For humans, understanding language is so natural. But training computers to perform these tasks is extremely challenging task and has huge implications in many areas of our lives. Automatic speech recognition (ASR) and natural language processing (NLP) based intelligent system can be used in several human machine interface applications both in consumer and industrial sector. A discussion on describing the architecture, building blocks, performance and applications for such system that would use latest ASR and NLP APIs is covered.

## KEYWORDS

i523, HID333, HID337, Natural Language Processing, Automatic Speech Recognition, Voic Recognition

## 1 INTRODUCTION

The advancements of digital signal processing, large data processing and natural language processing technologies made speech/voice recognition applications more sophisticated to help solving social and industrial problems. For example, having an intelligent automatic voice recognition system in home to recognize and differentiate between the family members and outsiders would add great value to modern society in terms of assisting in their busy life as well provide necessary help/guidance in offering day-to-day problem solutions, personalized entertainment, and safety/security. In another example, these systems can provide personalized customer care experience through voice and face recognition by engaging them based on their interests/hobbies. Google home, Alexa, and Siri have become part of the main stream human life activities to seek information and get entertainment by directly speaking with these devices.

The hypothetical intelligent voice system would need the following technologies to work together:

- Highly efficient voice sensors and high speed digital signal processors
- Automatic Voice Recognition (AVR) hardware and software algorithms
- Machine Learning (ML) algorithms to classify and learn the voice patterns

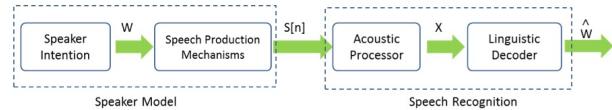


Figure 1: Conceptual Model of Speech Production

- Machine learning algorithms to understand family members habits and behaviors
- Natural Language Processing (NLP) algorithms to precisely recognize and process the voice data

Open Source and/or Other tools:

- Google Cloud Speech API or Alexa Voice Service (AVS)
- Audio processing hardware and software algorithms
- Natural language processing (NLP) to analyze the speech of family members, friends and strangers
- Interfacing with email servers, phone, text message servers

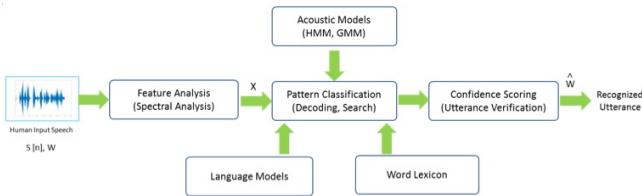
The following sections are organized to review some of the technologies available in the industry and universities and discuss the potential application concepts for home and industry.

The sections are broadly discussed on speech recognition and speaker recognition. In the speech recognition the focus is on detecting the words irrespective of the personal differences whereas speaker recognition is focused on detecting the physical speaker by discarding the words and their meanings. In other words, Speech recognition represents speech content and disregards speaker identity whereas speaker recognition represents speaker identity but disregards speech content.

A simple conceptual model of human speech production and recognition is shown in Figure 1. The human speech waveform creation process from the speaker intention is referred as Speaker Model which reflects speaker's accent and choice of words. The Speech Recognizer consists of acoustic processor which analyzes the speech signal and converts it into a set of acoustic (spectral, temporal) features followed by linguistic decoder to estimate the words of spoken sentence.

## 2 SPEAKER RECOGNITION THEORY

The speaker recognition technology is multidisciplinary, which requires hardware based sensors to convert voice in to electrical signals, speech processing module that converters the electrical



**Figure 2: Speaker Recognition Concept**

signals in to digitized data using advanced digital signal processors (DSP). The basic recognition process involves modeling the acoustic data and the natural language to search for patterns. Figure 2 shows the basic concept of voice recognition.

As speech is a sound pressure wave, its conversion in to electrical signal and then in to digital signal introduces distortion. As shown in the figure, acoustic front-end requires several signal processing components such as spectral shaping, spectral analysis, spectral modeling, and parametric transformation. These components will condition the signal and establish the spectral measurements and parameters for acoustic modeling.

Once robust parameterization of speech signal is established, then spectral dynamics or changes of the spectrum with respect to time will be captured. Typically, speaker recognition can be achieved by using differentiation of spectral features, which requires temporal derivatives of the voice spectrum. These temporal derivatives are commonly approximated by differentiating cepstral features using a linear regression.

## 2.1 Feature Analysis

There are no standard set of features for speech recognition. Instead, various combinations of acoustic, articulatory, and auditory features have been utilized in a range of speech recognition systems [9]. The input human speech signal,  $s[n]$ , is converted to the series of feature vectors,  $X = [x_1, x_2, \dots, x_T]$ , by the feature analysis or spectral analysis module. These feature vectors represent the temporal spectral characteristics of the speech signal in the form of mel frequency cepstrum coefficients.

## 2.2 Pattern Classification

The pattern classification is the process of grouping the patterns, which are sharing the same set of properties [2]. The pattern classification involves in computing a match score in speaker recognition system. The term match score refers the similarity of the input feature vectors to some model. Speaker models are built from the feature vectors extracted from the speech signals. Based on the feature extraction a model of the voice is generated and stored in the speaker recognition system. There three major techniques Dynamic Time Warping (DTW), Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) typically used in pattern classification process.

## 2.3 Speaker Acoustic Models

In the process of recognizing speaker voice, the speaker model will be created and trained with acoustic characteristics of the voice. The typical speaker recognition process involves two specific tasks: verification and identification. In the verification, the goal is to determine from a voice sample if a person is who he or she claims. In the identification, the goal is to determine which one of a group of known voices best matches the input voice sample. In either task the speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). The success in both tasks depends on extracting and modeling the speaker-dependent characteristics of the speech signal which can effectively distinguish one talk from another.

A brief description of some of the speaker modeling methods typically used is given below.

**Nearest Neighbor:** In this technique, feature vectors from enrollment (training) speech are retained to represent the speaker. During the verification, the match score is computed as the accumulated distance of each test feature vector to its  $k$  nearest neighbors in the speaker's training vectors.

**Neural Networks:** These models are trained to discriminate between the speaker being modeled and some alternate speakers.

**Hidden Markov Models (HMM):** The temporal evolution of speech signal features/characteristics and model statistical variations of the features are encoded to provide statistical representation of speaker.

**Template Matching:** In this method, the model contains a template with sequence of feature vectors. During the verification a match score is produced by using dynamic time warping (DTW) to align and measure the similarity between test phrase and speaker template.

## 3 NATURAL LANGUAGE PROCESSING (NLP)

NLP is the process of making machines to understand and interpret human languages just the way human beings understands. Speech recognition is the process of analyzing the acoustic data to extract the speech content. In this process speech will be converted in to text as first step. Then, the converted text will be fed to Natural Language Processing (NLP) algorithms for extracting the words and meaning. Machine learning algorithms are used in conjunction with language models to recognize text in natural language processing systems, which may also employ speech models and hardware/software specialized to process and recognize speech. Though human language is ambiguous and unstructured to be interpreted by computers, with the help of NLP, this huge unstructured data can be analyzed for finding the meaning contained inside the data.

Analyzing language for its meaning is a complex task. Modern speech recognition research began in the late 1950s with the advent of the digital computer [4]. The 1960s saw advances in the automatic segmentation of speech into units of linguistic relevance like phonemes, syllables [5]. And now with advancements in the field of Artificial Intelligence, deep machine learning algorithms have been used in many aspects of speech recognition like Part-of-speech Tagging, Word tokenization, Intent Extraction, phoneme

classification, and speaker adaptation. In the context of Speech Recognition, NLP involves 4 basic steps.

**Morphological Analysis:** Morphological analysis is the identification, analysis, and description of the structure of a given language's root words, word boundaries, affixes, parts of speech, etc. Non word tokens like punctuation are separated from the words. The term Morpheme means the "minimal unit of meaning". For example: In the word "unhappiness" there are three morphemes "un", "happy", "ness" each carrying its own meaning. Morphology treats with the different conjugations of a word and the forms it can take. For example word "sheep" can be singular or plural; a verb can have different tenses.

**Syntactic Analysis:** Syntactic analysis is the process of analyzing a linear sequence of words in natural language adhering to the rules of a formal grammar of the language. Processing a sentence syntactically includes determining the subject, predicate, verbs, adjectives, pronouns, etc. In this phase, linear sequence of words are converted into hierarchical tree structures that explains how words associate to each other. All most all Syntactic analysis procedures have two components:

- **Grammar:** A declarative expression of syntactic features about the language. It is the specification of the legal structures of a language. It constitutes 3 basic components: Terminal Symbols, Non Terminal Symbols and Rules (productions).
- **Parser:** Algorithm that compares the grammar against the input sentences to produce parsed structures called Parse Trees. Parsing can be done in two ways:
  - **Top-Down Parsing:** Begin with the start symbol and apply the grammar rules forward until the symbols at the terminals of the tree corresponds to the components of the sentence being parsed [1].
  - **Bottom-Up Parsing:** Begin with the sentence to be parsed and apply the grammar rules backward until a single tree whose terminals are the words of the sentence and whose top node is the start symbol has been produced [1].

**Semantic Analysis:** Semantic analysis is the process of linking syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. In this phase, the structures created by Syntactic analysis are assigned meanings.

**Pragmatic Analysis:** Pragmatic Analysis is how sentences are used in different situations and how use affects the interpretation of the sentence. Means what was said is reinterpreted as what it actually means. For example: the sentence "What is the time now?" should be interpreted as request instead of Question.

### 3.1 NLP Techniques

NLP techniques are broadly categorized into *Rule based (human driven)* and *Statistical based (data driven)*:

**Rule based:** Rule based (human driven) approach requires huge human effort. Grammars and semantic components are prepared in the form of many carefully handcrafted rules by highly skilled linguists. Rule based approaches takes time, money and trained personnel to make and test the rules. Also rule engineering may not scale very well. To make Rule based approach more accurate, it requires large number of complex hand-written rules which is much more difficult and laborious task. After certain number of rules, addition of any more rules going to increase the complexity of systems and makes the systems unmanageable. Once there are hundreds of rules, they start interacting in complex ways and becomes difficult while updating or adding any new rules. Rule based approaches have very poor generalization, but for the languages with fewest speakers, rules-based is the best approach since there exist not enough large corpora to go for Statistical approach. The best known parser with a rule base backbone is the RASP(Robust Accurate Statistical Parsing) system that combines rule-based grammar with a probabilistic parse selection model [7].

**Statistical based:** Statistical (data driven) approaches treats natural language processing as a *machine learning* problem. They use supervised or unsupervised statistical machine learning algorithms. This method applies learning algorithm to a large body of previously translated text (large data) known as a parallel corpus. Systems based on Statistical approach can be made more accurate by simply supplying more input data. The main advantage of the statistical approach is its language Independence. Provided there are annotated data, the same algorithm can be used for learning rules or models for any language. The statistical approach is significantly leading in terms of accuracy against manually annotated corpora, as well as in overall number of statistical parsers compared to the number of rule-based parsers. Fast, cheap computing hardware, advances in processor speed, random access memory size, secondary storage, and grid computing making Statistical approach as popular choice. One example parser with his approach is Malt-Parser, a data-driven parser-generator for dependency parsing that supports several parsing algorithms and learning algorithms and allows user-defined feature models, consisting of arbitrary combinations of lexical features, part-of-speech features and dependency features. The most significant disadvantage of this approach is the requirement of large amounts of training data in the form of large NL text corpora.

Efficient approach is to use both approaches, first use a rule-based model, then use its results as data for the statistical model.

### 3.2 Common usage of Deep learning algorithms in NLP

#### Newural Networks:

- Part-of-speech Tagging
- Word tokenization
- Named Entity Recognition
- Intent Extraction

#### Recurrent Neural Networks:

- Machine Translation

- Question Answering System
- Image Captioning

#### **Recursive Neural Networks:**

- Parsing Sentences
- Sentiment Analysis
- Relation Classification

#### **Convolutional Neural Networks:**

- Sentence/Text Classification
- Relation Extraction and Classification
- Semantic Relation Classification

### **3.3 Speech Recognition Technologies**

**Google Cloud Speech API:** Google Cloud Speech API [8] converts audio to text by applying powerful neural network models in an easy to use API. The API recognizes over 110 languages and supports audio files up to three hours in length. Two basic use cases where Google Cloud Speech apis can be best applied are

- human-computer interaction
- speech analytics on human-to-human interactions.

**IBM Watson Speech to Text:** Powerful real-time speech recognition software. Automatically transcribe audio from 7 languages in real-time. Rapidly identify and transcribe what is being discussed, even from lower quality audio, across a variety of audio formats and programming interfaces [6].

**Dragon NaturallySpeaking:** Dragon NaturallySpeaking (DNS) [3] is a speech recognition software package developed by Dragon Systems of Newton, Massachusetts. It recognizes and transcribes words at a high speed, and gives flexibility to dictate for any situation.

Carnegie Mellon University's Sphinx toolkit, HTK toolkit((free but copyrighted)) and Kaldi tookkits are some good software resources for speech recognition development.

### **4 CONFIDENCE SCORING OR SPEAKER VERIFICATION**

The confidence scoring process is used to provide a confidence score for each individual word in the recognized string. The scores are produced by extracting confidence features from the computation of the recognition hypothesis and processing the features using accept/reject classifier for word utterance hypothesis. The output of the confidence classifiers can then be incorporated into the parsing mechanism of the language understanding component [10].

### **5 BIG DATA CONTEXT IN SPEECH ANALYTICS**

To get better insight in to customer behavior, satisfaction, and trends information businesses are depending on big data technologies for voice analysis by analyzing large volumes of call data. The use of voice analytics combined with big data technologies will help call centers to improve performance by reducing the call time and repeat calls, providing customer satisfaction information etc. When applications need continuous recording and processing of large volumes of human speech data for home, industry or public enterprise security/information/entertainment purposes then big

data technologies will help meeting the computing and storage demands.

### **6 INTEGRATED SPEECH AND VOICE RECOGNITION APPLICATIONS**

Voice biometrics, customer service, truth detection, and personal voice assistant are some of the applications currently being used by the industry with speech recognition and analytics as key underlying technologies. Voice recognition technology has been in use by security systems with voice activated locks, law enforcement and criminology for truth detection.

In the customer service industry, speech analytics is playing key role to get complete insight in to customer behaviors and interests. Customers will interact with service providers using multiple channels such as email, social media, SMS, phone call, and in-person etc. The technology advancements are creating even more channels or options to interact with customers and service providers. However, with speech analytics systems the business can get more hidden insights for improving the customer satisfaction and loyalty. The capabilities such as phonetic-indexing, speech-to-text transcripts, speaker separation, emotion detection, and hot topics etc. are already in use by several businesses to improve their customer service performance.

The integrated speech and voice recognition will take the solution use cases one step beyond the current applications use and help improving the business performances. For example, systems will recognize returning customers with voice recognition technology and engage them with personalized interests/conversations.

### **7 CONCLUSION**

The voice, speech recognition technologies and NLP combined with big data technologies can be used in solving much complex problems than the current applications. Potential applications include personalized customer services, personal voice assists, and public information desks etc.

An attempt has been made to explain speech and speaker recognition concepts along with big data technology use in the applications. Then, some of the typical speaker acoustic models and pattern classification and platter recognition methods have been listed. There are several companies and entrepreneurs researching to create better Natural Language Processing (NLP) solutions and teach computers how to better understand human communication. Some of the unsolved technologies such as cross language translators and accurate speaker recognition are still in research, which when solved can unleash the great potential across the world.

### **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] R C Chakraborty. 2015. Natural Language Processing. (2015). [http://www.myreaders.info/10\\_Natural\\_Language\\_Processing.pdf](http://www.myreaders.info/10_Natural_Language_Processing.pdf)
- [2] Dr E. Chandra and K. Manikandan M. S. Kalaiavani. 2014. A Study on Speaker Recognition System and Pattern classification Techniques. (2014). <https://www.ijsreeice.com/upload/2014/february/IJSREEICE1H...a...kALAI...A.study.pdf>

- [3] Nuance Communications. 2017. (2017). <https://www.nuance.com/dragon.html>
- [4] Jacqueline R. Dalton and Cindee Q. Peterson. 1997. The Use of Voice Recognition as a Control Interface for Word Processing. *Occupational Therapy In Health Care* 11, 1 (1997), 75–81. [https://doi.org/10.1080/J003v11n01\\_05](https://doi.org/10.1080/J003v11n01_05)
- [5] Daryl H. Graf and Richard D. Peacocke. 1990. An Introduction to Speech and Speaker Recognition. *Computer* 23 (1990), 26–33.
- [6] IBM. 2017. (2017). <https://www.ibm.com/watson/services/speech-to-text/>
- [7] Vojtch KOVff. 2014. Automatic Syntactic Analysis for Real-World Applications [online]. (2014). <https://nlp.fi.muni.cz/~xkovar3/thesis.pdf>
- [8] Google LLC. 2017. (2017). <https://cloud.google.com/speech/>
- [9] Lawrence R. Rabiner and Ronald W. Schafer. 2007. Introduction to Digital Speech Processing. *Found. Trends Signal Process.* 1, 1 (Jan. 2007), 1–194. <https://doi.org/10.1561/2000000001>
- [10] J. Hazen Timothy, Burianek Theresa, Polifroni Joseph, and Seneff Stephanie. 2000. Recognition Confidence Scoring for Use in Speech Understanding Systems. (08 2000), 49–67 pages. [https://www.researchgate.net/publication/2645827\\_Recognition\\_Confidence\\_Scoring\\_for\\_Use\\_in\\_Speech\\_Understanding\\_Systems](https://www.researchgate.net/publication/2645827_Recognition_Confidence_Scoring_for_Use_in_Speech_Understanding_Systems)

## A WORK BREAKDOWN

### A.1 HID 333:Anil Ravi

- Identified Paper2 topic
- Created Paper2 draft sections
- Finalized speech recognition theory
- Reviewed all sections of the paper

### A.2 HID 337:Ashok Reddy Singam

- Worked on NLP and its subsections
- Editing Latex template using ShareLatex online tool
- Managed JabRef entries
- Reviewed the draft paper

## bibtex report

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

## Database file #1: report.bib

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3131 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3229 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3229 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3229 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3229 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea while executing---line 3229 of file ACM-Reference-Format.bst

```
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea
while executing---line 3229 of file ACM-Reference-Format.bst
(There were 20 error messages)
make[2]: *** [bibtex] Error 2
```

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

```
[2017-11-06 17.38.31] pdflatex report.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.17 (TeX Live 2016) (preloaded format=pdflat
Missing character: ""
Typesetting of "report.tex" completed in 1.2s.
```

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

```
name: Ashok Reddy Singam
hid: 337
paper1: Nov 01 17 100%
paper2: Nov 06 17 100%
project: not started
```

```
yamlcheck
-----
```

```
wordcount
-----
```

```
5
wc 337 paper2 5 3109 report.tex
wc 337 paper2 5 3096 report.pdf
wc 337 paper2 5 393 report.bib
```

```
find "
```

```
-----
```

```
passed: True
```

```
find footnote
-----
```

---

```
passed: True
```

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

---

```
passed: False
```

```
floats
```

---

```
64: \begin{figure}
65: \includegraphics[width=1.0\columnwidth]{images/speechrecognition.p
  df}
74: \begin{figure}
75: \includegraphics[width=1.0\columnwidth]{images/speakerrecognition.
  pdf}
```

```
figures 2
tables 0
includegraphics 2
labels 0
refs 0
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 referred to: (refs >= labels)
```

```
Label/ref check
```

```
68: A simple conceptual model of human speech production and
recognition is shown in Figure 1. The human speech waveform
creation process from the speaker intention is referred as Speaker
Model which reflects speaker's accent and choice of words. The
Speech Recognizer consists of acoustic processor which analyzes
the speech signal and converts it into a set of acoustic
(spectral, temporal) features followed by linguistic decoder to
estimate the words of spoken sentence.
71: The speaker recognition technology is multidisciplinary, which
requires hardware based sensors to convert voice in to electrical
signals, speech processing module that converters the electrical
signals in to digitized data using advanced digital signal
processors (DSP). The basic recognition process involves modeling
the acoustic data and the natural language to search for patterns.
```

Figure 2 shows the basic concept of voice recognition.  
passed: False -> labels or refs used wrong

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

find textwidth

---

passed: True

below\_check

---

bibtex

---

label errors

bibtex errors

This is BibTeX, Version 0.99d (TeX Live 2016)

The top-level auxiliary file: report.aux

The style file: ACM-Reference-Format.bst

Database file #1: report.bib

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Richard D. while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Daryl H. while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Richard D. while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Daryl H. while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Richard D. while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Daryl H. while executing---line 3085 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Richard D. while executing---line 3085 of file ACM-Reference-Format.bst

Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Peacocke, Daryl H. while executing---line 3131 of file ACM-Reference-Format.bst

Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3131 of file ACM-Reference-Format.bst  
Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 1 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
Name 2 in "Daryl H. Graf, and Richard D. Peacocke," has a comma at the end for entry Pea  
while executing---line 3229 of file ACM-Reference-Format.bst  
(There were 20 error messages)

bibtex\_empty\_fields

---

entries in general should not be empty in bibtex

find ""

---

passed: True

---

ascii

---

non ascii found 225  
non ascii found 345

---

The following tests are optional

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

find newline  
-----

73: \newline

183: \newline

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

find cite {  
-----

passed: True