Use Cases in Big Data Software and Analytics

Vol. 2, Fall 2017

Bloomington, Indiana

Monday 11th December, 2017, 06:08

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

Contents

Chapter 1

Preface

1.1 Disclaimer

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

1.2 Citation

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

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

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

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

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

1.3 List of Papers

HID	Author	Title
104, 216	Jones, Gabriel, Millard, Mathew	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
107	Ni,Juan	NoSQL Databases in support of Big Data Applications and Analytics
109	Shiqi Shen	Big Data and Application in Amazon
201	Arnav, Arnav	Using MQTT for Communication in IoT Applications
203	Chandwani, Nisha	Big Data Analytics Using Regression Techniques
204	Chaturvedi, Dhawal	Big Data and Support Vector Machines
208	Jyothi Pranavi Devineni	Algorithms for Big Data Analysis
209	Han, Wenxuan	Clustering Algorithms in Big Data Analysis
212, 225, 210	Kumar, Saurabh; Schwartzer, Matthew; Hotz, Nicholas	Can Blockchain Adoption Mitigate the Opioid Crisis Through More Secure Drug Distribution?
211	Khamkar, Ajinkya	Machine learning optimizations for big data
212, 225, 210	Kumar, Saurabh; Schwartzer, Matthew; Hotz, Nicholas	Can Blockchain Adoption Mitigate the Opioid Crisis Through More Secure Drug Distribution?
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, 104	Millard, Mathew, Jones, Gabriel	Big Data = Big Bias? The Fallibility of Big Data
218	Niu, Geng	How Big Data Transform Education
219	Syam Sundar Herle Parampali Sreenath	Big Data and Sentiment Analysis
224	Rawat, Neha	Big Data Applications in the Energy and Utilities Sector

212, 225, 210	Kumar, Saurabh, Schwartzer, Matthew, Hotz, Nicholas	Can Blockchain Mitigate the Opioid Crisis Through More Secure Drug Distribution?
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 Spark
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	Big Data and Adaptive Learning
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, 323	Andres Castro Benavides, Uma Kugan	Big Data Security and Privacy.
306	Murali Cheruvu	Why Deep Learning matters in IoT Data Analytics?
309	Dubey, Lokesh	BigData Applications in Social Media for Marketing
310	Kevin Duffy	Big Data Applications for Vehicle Crash Prediction
312	Neil Eliason	Big Data Applications in Historical Studies
313	Tiffany Fabianac	Big Data Applications in Laboratories
315	Garner, Jeffry	Concussions and Big Data's Opportunities and Challenges
316	Robert Gasiewicz	Big Data on IoT Smart Refrigerators
319	Mani Kumar Kagita	Mini Project: ESP8266 and Raspberry Pi Robot Car
320	Elena Kirzhner	Overview of Python Data Visualization Tools
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
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, 333	Ashok Reddy Singam, Anil Ravi	Natural Language Processing (NLP) to Analyze Human Speech Data
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
337, 333	Ashok Reddy Singam, Anil Ravi	Natural Language Processing (NLP) to Analyze Human Speech Data
338	Anand Sriramulu	A comparative study of Kubernetes and Docker Swarm and Advantages of Singularity Container to HPC World

340	Timothy A. Thompson	BigchainDB: A Big Database for the Blockchain?
341	Tibenkana, Jacob	Big Data Applications for Visualizations
343	Borga Edionse Usifo	Big Data Applications and Manufacturing
345	Ross Wood	How the Datafication of Activity is Improving Human Health
346	Zachary Meier	Netflix use of Big Data Visualization
347	Jeramy Townsley	Sociological Qualitative Methods Using Big Data
348	Budhaditya Roy	Security aspect of NOSQL database in Big Data Applications

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 Frad in real life.

KEYWORDS

i423, hid106, Data Science, Big Data Analytics, Cloud Computing,frad 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 [5]. 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 [1]. What's more, the trends in big data (the evolution in unstructured information) are prone to numerous fraud, which can go without notice is 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 [4]. 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%) [3]. 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 [2, 4]. 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 [1]. 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 [5]. 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 [3]. 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 one-portion 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 [4] . 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 [3]. 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 [5]. 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[1, 3]. 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 [1, 5]. 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 [1]. 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

[1] Chui M. Brown, B. and J Manyika. 2011. Are you ready for the era of fibig datafi? *McKinsey Quarterly* 4, 1 (2011), 24–35.

[2] Chiang R. H. L. & Storey V. C Chen, H. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012).

[3] A. A. Crdenas, P. K. Manadhata, and S. P. Rajan. 2013. Big Data Analytics for Security. *IEEE Security Privacy* 11, 6 (2013), 74–76.

[4] Shaun Hipgrave. 2013. Smarter fraud investigations with big data analytics. *Network Security* 2013, 12 (2013), 7–9.

[5] Eric Siegel. 2013. *Predictive analytics: The power to predict who will click, buy, lie, or die.* Number 103-110. John Wiley & John Wiley & Sons.

Can Blockchain Mitigate the Opioid Crisis Through More Secure Drug Distribution?

Saurabh Kumar
Indiana University
Bloomington, IN 47408, USA
kumarsau@iu.edu

Mathew Schwartzer
Indiana University
Bloomington, IN 47408, USA
mabschwa@iu.edu

Nicholas J Hotz
Indiana University
Bloomington, IN 47408, USA
nhotz@iu.edu

ABSTRACT

Like TCP/IP in the 1970s and 1980s, blockchain is a new, intriguing but grossly misunderstood technology that is still in its infancy. It is commonly misunderstood as just a technology for Bitcoin and cryptocurrencies. However, blockchain's use cases extend beyond just financial transactions and cryptocurrencies and have the potential to transform nearly every industry including healthcare and supply chain. As the technology matures, additional transformative use cases could expand into drug distribution; specifically, a blockchain system for prescription opioid distribution could theoretically mitigate some aspects of the opioid crisis.

KEYWORDS

HID 210, HID 212, HID 225, i523, Blockchain, Opioid Epidemic, Pharmaceutical Supply Chain, Healthcare

1 INTRODUCTION

1.1 The Need to Modernize Global Record Keeping

Contracts, transactional records, and verification systems are part of the foundational core of the global economy. However, as Iansiti and Lakhani [25] explain, these tools have not modernized to keep up with the needs of the rapidly evolving global economy and are "like a rush-hour gridlock trapping a Formula 1 car." Records and transactions are still being managed as they were in the 20th century which creates broad consequences for nearly every industry including supply chain and healthcare.

In supply chain, data management methods for records and logistics are usually inconsistent across the different levels of a supply chain [3]. The outdated record management method encourages redundant data to be stored at the same organization as well as across the supply chain which increases IT maintenance costs and decreases trust and transparency [18]. These issues prevent a tertiary party, like the government, to effectively scrutinize records.

Outdated data management processes also negatively impact healthcare. In the USA in 2014 healthcare fraud cost an estimated $272 billion [12], and in 2016, healthcare data breaches impacted over 27 million patients [9]. Today, medical data management is stifled by antiquated technology that limits patients' ability to manage and control access to their electronic medical records [13]. In addition, pharmaceutical supply chains are enervated by current record-keeping technologies. Transactional records are rarely shared across pharmaceutical supply chain organizations which consequently increases inventory levels [35]. As a result, total healthcare cost and the opportunity for counterfeit drugs increases [48]. In addition, verification systems are often independent among supply chain retailers and prescribers. The lack coordination opens the door for "doctor shopping" and greater prescription medication abuse [14].

1.2 Rational Exuberance for Blockchain

Blockchain, "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way" [25], has the potential to resolve these and other fundamental problems of the global economy by overcoming many of the antiquated shortcomings of the traditional means of managing and verifying contracts and transactions. However, like TCP/IP in the 1970s and 1980s, blockchain is an immature technology that faces numerous challenges to mass adoption. In spite of its current limitations, blockchain is already seeing promising applications in various industries extending beyond just finance including healthcare and supply chain. One particularly exciting use case sits at the intersection of healthcare and supply chain for a more secure distribution system for opioid medications that could potentially mitigate the opioid crisis.

2 BLOCKCHAIN OVERVIEW

2.1 The Blockchain Framework

Blockchain is a foundational technology comprised of numerous technological processes and entities. Some of the most significant pieces follow.

2.1.1 Node. Nodes are the individual units connected to the blockchain network. They are computers with adequate software to maintain a blockchain. The blockchain network connects all the nodes and can read and write data to a block [47] [28].

2.1.2 Block. Blocks are the group of records, bundled together by nodes. They follow a specific set of rules and have limited size. Blocks are also linked to the last generated block, thus forming a chain [47].

2.1.3 Smart Contracts. Smart contracts are the codes with time stamps to represent a contract [47]. Iansiti and Lakhani [25] believe that "'smart contracts' may be the most transformative blockchain application at the moment," because they allow for automatic payments whenever contract conditions are met.

2.1.4 Submit Transaction. In case of a new transaction submission to the network, an individual node circulates it to all the other nodes in the network [47]. The main purpose of circulation is approval.

2.1.5 Transaction Approval. When a transaction is submitted and circulated in the network, each node verifies it. Invalid transactions are deleted [47].

2.1.6 Consensus. For multiple systems to work in a distributed network, they must have an agreement. Such a structure is useful in case of fault tolerance when those agreed set of protocols help to restore the data [47].

2.2 Data Types in Blockchain

There are three major types of data stored on a blockchain, namely un-encrypted, encrypted and hashed [18].

2.2.1 Un-encrypted Data. All the organizations have read access to the un-encrypted data. Such data is fully transparent and facilitates immediate dispute resolution [18].

2.2.2 Encrypted Data. The encrypted data can only be read by the organizations with the access to such data. This means an organization should have a decryption key to read to read the encrypted data. Encrypted data provides restricted access but is also stored in every node in the blockchain. In case of a dispute, the decryption key could be used by different organizations to rectify the entry or deletion of any record [18].

2.2.3 Hash Data. Hash data is also a hidden data, where hash keys act like fingerprints to represent changes or entry for any data record. Each organization can easily confirm their hash keys. Breaking the hash key is nearly impossible. Only the hash key is in the blockchain while the record data is stored off-chain by individual organizations. Data could be revealed, in case of a dispute, by the respective organization [18].

2.3 Benefits of Blockchain

Blockchain's framework and data types provide such broad-ranging benefits that blockchain has been proposed as the "cure" to solve many of the world's problems. This exuberance stems from the fundamental benefits that are cornerstones to nearly every industry.

2.3.1 Trust. Blockchains enable parties that do not know each other to trust each other. No single organization is trusted to maintain the records. Instead, all organizations must approve the contents of the record in order to avoid disputes. Therefore, records should have a time stamp and an origin proof. Normally, a third party facilitates this requirement. Blockchains can provide an alternate solution, where organizations jointly manage the records and preventing corruption by a single organization [18].

2.3.2 Access. Blockchains allow for greater control over what information is and is not accessible. The technology enforces identical data to be stored by each organization. When one copy is updated, all the other copies are also updated. This eliminates the need for a third party to facilitate management of records [26]. Alternatively, different levels of read and write access could be provided to different organizations. Although some meta data should be stored in the public ledger.

2.3.3 Redundancy and Security. Blockchain also assists in providing security by disallowing redundancy at the same node. In the areas of logistics and inventory data, blockchain provides a new approach to supply chain management. The core logic of blockchain does not allow duplicate entries to be created in the same place [3]. A unique inventory can have a single entry with multiple updates, but not duplication. This prevents the organizations from creating false information. In the example of a drug inventory, the shipment status for a batch of drugs will be updated for everyone, everywhere. Each entry could be tracked back to its origin [3].

2.3.4 Transparency. Transparency in a business helps to grow trust among organizations. Sharing information can improve relationships among these organizations. Without blockchain, transparency is hard to achieve. Blockchains can help improve the visibility of contracts, legal documents as well as other inter-organization data [47]. Organizations are not obligated to show all of their data. Some access can be provided to data that could be useful to other organizations and a shared collection of records can also be stored and managed by co-operation from different organizations.

2.3.5 Low Transaction Costs. Through by-passing third-party verification systems such as brokers, lawyers, or banks, blockchain could significantly reduce transaction costs. Not only will this lower costs for existing transactions, it could open up the market for micro-payments [25].

2.4 Challenges to Blockchain Mass Adoption

While it indeed has the potential to help a wide variety of the world's problems, it should not be viewed as a panacea. Blockchain is not mature enough to support mass-market adoption and faces numerous challenges. Rabah [42] states that to be effective, blockchain needs to overcome its shortcomings of lacking standard protocols, unclear regulation, large energy and computing power consumption, privacy, cultural adoption, and high initial capital requirements. Tapscott and Tapscott [51] agree that its current technical infrastructure is not sufficient, its energy consumption and computational requirements are not sustainable, and user-friendly systems have yet to be designed that would allow for mass market adoption.

Society would have to dismantle many technological, governance, organizational, and cultural barriers to create new foundations for a new world economy that relies heavily on blockchain [25]. This will come at the cost of some existing societal norms, core business functions, and people's jobs [25] [42].

2.5 Technology Adoption Lifecycle

Iansiti and Lakhani [25] argue that the process for mass adoption of blockchain may take longer than expected but will follow a fairly predictable technology adoption pattern that parallels the adoption of TCP/IP (transmission control protocol / internet protocol). TCP/IP started as *single-use* and matured to *localized uses*, *substitutions*, and *transformations*. It was introduced as a *single-use* in 1972 for e-mail in ARPAnet, a precursor to commercial internet for the US Department of Defense. Met with skepticism, this technology slowly gained traction among some firms in the 1980s and early 1990s for *localized use* and did not become mainstream until the emergence of World Wide Web in the mid-1990s. This then paved the road for infrastructure companies to provide the necessary hardware and software to establish "plumbing" systems for the internet.

Once the technical infrastructure was mature enough, companies then developed businesses that *substituted* existing services with online services (such as Amazon books instead of Borders). Finally, a wave of companies created *transformative* applications that fundamentally changed service experiences (such as Napster in the music industry or Skype in telecommunications).

Similarly, blockchain was also launched for a *single use* in 2009 for Bitcoin, a virtual currency. Blockchain has matured to extend beyond cryptocurrencies and is now being applied for various *localized uses* including in healthcare and supply chain. It took over 30 years for TCP/IP to realize its potential, and blockchain will likewise require decades to mature into a revolutionary economic force. However, companies can start planning for this revolution today and implement blockchains that follow seven design principles [25] [51].

2.6 Seven Design Principles for Blockchain

Tapscott and Tapscott [51] in their book *Blockchain Revolution* propose seven design principles that, when appropriately applied, can help blockchain move down the technology adoption lifecycle and create more honest, cost-effective, and accountable systems.

2.6.1 Networked integrity. Because all organizations on the blockchain must approve updates, "Participants can exchange value directly with the expectation that the other party will act with integrity." [51].

2.6.2 Distributed Power. Since the blockchain is distributed across a broad network, it cannot be dismantled by authoritarian power, hackers, or other bad actors. There are no single points of failure and the blockchain can still perpetuate even if numerous nodes are compromised [51].

2.6.3 Value as Incentive. Blockchains can align incentives of individual participants with the interests of the entire blockchain. This minimizes organization problems and conflicts of interests [51].

2.6.4 Security. Blockchains can protect against hackers, malware, ransomware, and identity theft by using a variety of security features. Public key infrastructures, private keys, public keys, and verification methods verify participant activities and prevent bad actors from overriding the network [51].

2.6.5 Privacy. Blockchains can and should provide participants with the freedom to expose as little or as much information about themselves as they desire. This allows a participant to act anonymously when desired or to share sensitive information with only appropriate parties when needed [51].

2.6.6 Rights Preserved. To protect against counterfeit items, a blockchain can serve as a public ledger of ownership [51].

2.6.7 Inclusion. Currently, access to certain financial services is limited to those who are deemed "creditworthy". Blockchains can and should have significantly lower bars of entry that are not managed by banking institutions so that even a poor rural former on a remote corner of Earth who isn't creditworthy, could participate in the blockchain [51].

3 BLOCKCHAIN APPLICATIONS

3.1 Supply Chain

Blockchain, being a public ledger, can be used in different domains with slight variation in its core attributes. While the general implementation says that the data of a single block is public to all the nodes, different sets of access rights could be provided to different classes of users. Such implementation of blockchain could be applied to a supply chain network.

A supply chain requires the involvement of various parties helping each other. This is generally a one-to-one chain network. Often, each organization uses different technologies for record keeping. Record keeping could involve any information ranging from direct communications to logistics. Trust is an important issue between organizations. Most of the organizations in a supply chain keep individual records, which are not public to other organizations in the supply chain. Organizations share some information like contracts or notarized data. An efficient management of such shared data can be accomplished using a blockchain. The blockchain provides the ability to collect, record, and notarize different types of shared data [18].

Blockchain could also facilitate storing and maintaining logistics data. Such an application could be useful in the field of healthcare, where the government wants to monitor the supply of drugs. An ideal scenario for this would be to mitigate issues like the opioid crisis. Blockchain technology could simplify storage and management of trusted information. It could provide easy access of such critical public sector information to government organizations while providing data security [50]. Blocks comprise of the data records. When these blocks are added to the chain, they become immutable. This means they cannot be deleted or changed by a single organization [50]. A consensus has to be reached by a majority of the organizations for changing any record. Such a feature helps to maintain the security of the records by eliminating data corruption. Each block is verified and managed using some shared protocols. This process can be automated to allow ease of data entry. Two uses cases are for counterfeit detection and data analysis.

3.2 Healthcare

Representing over 17% of the United States' GDP, healthcare costs continue to soar [14]. More effective data management could address many of healthcare's fundamental issues, and according to a 2011 McKinsey report [33], more effective health data management could save $300 billion annually. Current innovations focus on placing patients at the center, privacy and access, completeness of information, and cost [14]. Three interesting applications of blockchain for healthcare are in claims adjudication, cyber security and healthcare IoT, and electronic medical records [9].

3.2.1 Claims Adjudication and Fraud Prevention. In 2014, the Economist estimated that the United States wasted $272 billion dollars on healthcare fraud [12]. Blockchain could not only minimize fraudulent billing; but, by automating claims adjudication and billing processes, obviate the need for administrative and transactional costs through third parties. Gem Health and Capital One are developing a blockchain-based solution for healthcare claims management [9].

3.2.2 Cyber Security and Healthcare IoT. In 2016, there were 450 reported health data breaches, impacting 27 million patients. Hacking and ransomware were responsible for 27% of these breaches. Each additional connected medical device serves as a potential entry point for bad actors. With an estimated 20-30 billion healthcare IoT devices by 2020, blockchain could secure these devices and protect confidential data. Telstra, IBM, and Tierion are three companies that are developing cyber security solutions for connected healthcare devices [9].

3.2.3 Electronic Medical Records. Beleaguered by stifled technology development, limited ownership control by patients, fragmented information systems, and risks of electronic protected health information hacking, electronic medical records have perhaps the most important use cases for blockchain [57]. Blockchain can provide interoperability of healthcare information, improved security, patient-centric control, and immutable records [9]. Three examples of blockchain-based EMRs include MedRec, Medicalchain, and the Estonian eHealth Foundation. First, by leveraging smart contracts on the Ethereum blockchain, MedRec is a prototype system that provides patients with "one-stop-shop access to their medical history" and shows promise to give ownership of health information back to the patients who can selectively share access through a modern API interface in a secure manner [13]. Second, Medicalchain is a permissioned blockchain distributed on networks of international healthcare providers that allow patients to transfer medical records across national borders [14]. Third, a data security company called Guardtime is using its Keyless Signature Infrastructure system in partnership with the Estonian eHealth Foundation to store Estonian health records on a blockchain.

4 THE OPIOID CRISIS

4.1 Addiction Risk

Since the late 1990s, pharmaceutical companies have downplayed the addictive risk of opioids [38]. However, the addictive nature of prescribed opioid painkillers increases the "potential for unforeseen adverse events for the patient, including overdose, experience of physiological dependence and subsequent withdrawal, addiction, and negative impacts on functioning" [54]. Patients with wholesome medical intentions often fall victim to the pills' addictive nature. Misuse and eventual abuse of prescribed opioid painkillers are common: 21%-29% of patients prescribed opioids for chronic pain misuse them while 7.8%-11.7% develop an addiction [54]. Moreover, an opioid addiction often serves as a gateway to other illegal drug use. With similar highs, prescription opioid addicts often transition to heroin, an illicit street-made opioid, since it is cheaper and easier to obtain. In fact, 4%-6% of patients using prescribed opioids develop a heroin addiction [38]. Whereas, 75% of heroin users began their opioid addiction with prescription opioids [7].

Despite these risks, opioids are still prescribed at alarming rates. In fact, the United States, with about 5% of the world's population, consumed 80% of the world's opioid prescriptions from 2001-2010 [54]. Between 1999 and 2015 the amount of prescribed opioids painkillers such as codeine, fentanyl, oxycodone, Demerol, and Vicodin quadrupled. In the same time period, opioid-related deaths also quadrupled.

4.2 Health Impact

The epidemic has become so severe that in October 2017 President Trump was forced to declare it "a national health emergency" [37]. With no signs of stopping, this epidemic is burgeoning across America killing nearly 91 people a day [45].

In 2015, 33,091 Americans died from an opioid overdose with rural white males at the greatest risk of an opioid overdose. White Americans (27,056) died the most, followed by black Americans (2,741), and Hispanic American (2,507). Generally the middle-aged population was most at risk with the following percent distributions by age group [17]:

- Aged 0-24: 10% of the opioid-related deaths
- 25-34: 26%
- 35-44: 23%
- 45-54: 23%
- 55+: 19%

Males die nearly twice as frequently from an opioid overdose, representing 65% deaths compared with 35% for females [17].

4.3 Financial Impact

The health impacts are the primary reason for concern, but the financial liability associated with the epidemic is also increasing. The estimated financial impact of the crisis grew from $55.7 billion in 2007 [2] to $78.5 billion in 2013 [15]. Of the total economic burden, roughly 25% or $20 billion is conveyed to the public sector [15]. Partitioned between workplace, healthcare, and criminal justice costs, the overall financial burden will continue to rise until a reversal in current trends.

Opioid drug makers are also exposed to significant financial and legal liabilities as lawsuits accusing pharmaceutical companies of deceptive marketing are commonplace. After a U.S. Justice Department probe in 2007, the maker of OxyContin pleaded guilty to federal charges and paid $634.5 million. In later cases, OxyContin maker Purdue Pharma LP settled two additional cases for a combined $43.5 million. Since then governments litigating the culpability of opioid drug makers include "South Carolina, Oklahoma, Mississippi, Ohio, Missouri and New Hampshire as well as cities and counties in California, Illinois, Ohio, Oregon, Tennessee and New York" [43]. In a suit filed in April 2017 against the three largest drug retailers in the USA - CVS, Walgreens, and Walmart - lawyers for plaintiffs Cherokee Nation claim that the "Defendants turned a blind eye to the problem of opioid diversion and profited from the sale of prescription opioids to the citizens of the Cherokee Nation in quantities that far exceeded the number of prescriptions that could reasonably have been used for legitimate medical purposes" [21].

4.4 Responses to Mitigate the Crisis

The private sector, government, and academia alike recognize the importance of solving this crisis and are implementing strategies to help mitigate the opioid crisis.

4.4.1 Private Sector. Drug retailers are taking immediate action. In September 2017, CVS pharmacy announced actions to limit patient supply of prescription opioids to seven days, to restrict the

strength of opioids dispensed for first time patients and to install 750 more in-store drug disposal kiosks [5] [19].

A longer-term private sector solution is through the use of radio frequency identification (RFID) technology as a method to improve supply chain security [52] [56]. RFID tracking tags are small microchips that are either printed, etched, stamped, or vapor-deposited onto product labels and are intended to replace barcodes. RFID can be read without direct line of sight and at distances up to 30 feet. Research shows that RFID tags have the potential to reduce costs, increase transparency, and identify counterfeit lots. RFID tags have many advantages over current barcode tracking methods. RFID tags can hold up to 32,000 alphanumeric characters compared to just 20 in a barcode. RFID tags have a much higher upfront cost but decrease total supply chain cost due to the timely process to scan each individual barcode. And unlike RFID tags, barcodes are susceptible to wear and tear and are easily replicated. RFID technology also has its flaws. In addition to the higher upfront cost, each tag costs between 5-10 US cents, signficantly higher than bar codes. Moreover, they are vulnerable to electromagnetic interference and poor manufacturing, are larger, and require a much larger IT infrastructure [52] [27]. From a security and transparency perspective, RFID technology is a good option to conform to The Drug Quality and Security Act [1].

4.4.2 Government. Through policy and politics, the federal government is attempting to find solutions to the epidemic. In the same address President Trump declared the opioid epidemic a national health crisis, he proposed "really tough, really big, really great advertising" [11]. Tom Price of the U.S. Department of Health and Human Services outlined a more detailed federal long-term plan including, "improving access to treatment and recovery services, promoting use of overdose-reversing drugs, strengthening our understanding of the epidemic through better public health surveillance, providing support for cutting edge research on pain and addiction, and advancing better practices for pain management"[41]. Additionally, President Trump's Commission on Combating Drug Addiction and the Opioid Crisis repeatedly mentions "data sharing" as a method to cope and limit the opioid crisis [37].

Multiple studies indicate that states with strong prescription drug monitoring programs (PDMPs) show a significant reduction in the number of opioid-related deaths [39] [40]. Evidence suggests that 72% of physicians were aware of their states' PDMPs in 2015, but only 52% used their services. Physicians noted difficulties understanding the data formats and retrieval systems as the main barriers to continual use of PDMPs [46]. As a result, low registration rates are common in the 49 states that offer some form PDMPs [20].

Increasing access to Naloxone, an opioid antagonist that rapidly reverses the opioid overdose damage, may be the most important immediate solution to reducing opioid-related deaths [20]. Between 1998 and 2014, 52,283 naloxone kits were distributed among the 30 states with naloxone distribution programs resulting in 26,453 overdose reversals [20]. 27 states have "third-party prescription" laws that allow physicians to prescribe Naloxone to family and friends of individuals with an opioid addiction [20]. To further reduce opioid-related deaths states must reduce malpractice liability for physicians prescribing Naloxone and make Naloxone available without a prescription [20].

In addition, states have started to pass legislation protecting Good Samaritans. As of 2014, 23 states had laws protecting cooperating bystanders, from low-level misdemeanors and drug possession. Without these laws, bystanders are subject to criminal charges and even murder if it is proven they supplied the deadly drugs. Consequently, these laws are necessary to encourage immediate life-saving calls to 911 [4] [20].

Other solutions states should consider is access to medical marijuana, as Pardo [39] found that states with legal medical marijuana dispensaries have lower opioid-related deaths.

4.4.3 Academia. Academic research is helping to propose effective solutions to the opioid crisis. For example, Indiana University announced plans to commit $50 million and 70 researchers to find solutions that lead to a decline in opioid-related deaths [44]. In a similar proposal, researchers at the Network for Public Health Law, Boston University, and Northeastern University proposed a four-step solution including "improving clinical decision making and access to evidence-based treatment, investing in comprehensive public health approaches, and re-focusing law enforcement response "[10].

5 AN OVERVIEW OF PHARMACEUTICAL SUPPLY CHAINS

5.1 Network Nodes

Forward facing supply chain activities occur before a customer purchase. In a pharmaceutical supply chain, forward facing nodes includes manufacturers, warehouses, distributors, and retailers. Reverse facing supply chain activities occur after the sale and include collecting, recycling, redistributing, and disposing of unwanted medications.

5.1.1 Primary Manufacturing. Produces the main active ingredient [49].

5.1.2 Secondary Manufacturing. Often at a different geographic location for tax and labor reasons, secondary manufacturers combine the active ingredients produced by primary manufacturers and adding excipient substances. Secondary manufacturers produce distribution ready SKU medications through one or more of the following processes: granulation, compression, coating, quality control, and packaging [49].

5.1.3 Market Warehouses and Distribution Centers. Due to the cost of setup and cleaning, it is common for primary manufacturers to produce a years' worth of active ingredients for a particular medication in one batch. This strategy creates a lot of excess finished and work-in-progress inventory that is stored in warehouse and distribution centers [49].

5.1.4 Wholesalers. Roughly 80% of demand flows through wholesalers. The industry is highly competitive and consolidated. The largest five wholesalers accounted for roughly 45% of industry revenue [49] [23].

5.1.5 Pharmacies and Hospitals. The last node on the pharmaceutical supply chain before medications are distributed at a patient level. Major retailers include pharmacies CVS, Walgreens, Walmart, and Rite Aid and hospital systems such as Community Health

Systems, Hospital Corporation of America, and Ascension Health [49].

5.1.6 Reverse Supply Chain. The reverse supply chain is often overlooked as a key component of the pharmaceutical supply chain network. Few people take their unwanted medications to proper collection sites. Instead, medications are discarded in the trash and sewage. In fact, in 2003 the world disposed of at least $760 million worth of prescription medications [24]. By 2014, this number ballooned to an estimated $5 billion [32]. The roughly 10 million unused and unexpired prescription medications could be recycled and reused, but instead improper disposal leads to dangerous compounds in sewage effluent, surface water, and even drinking water [24] [32]. Hua, Tang, and Wu [24] suggest a combination of government subsidies, penalties, and marketing to encourage drug makers to collect unwanted and expired medications.

5.2 Weaknesses

The nature of the current pharmaceutical production and supply chain system creates multiple weaknesses.

5.2.1 Lead Time. Lead times, the time it takes between manufacturing and end sale, can take up to 300 days [49]. As a result, high safety stocks are needed to react to future demand.

5.2.2 High Service Levels. The necessity for on-time pharmaceutical products forces retailers to maintain high service levels, the targeted rate of stock-outs. In many cases and especially in hospitals, patient health relies on having the right medication at the right time. A failure to meet this immediate demand could lead to fatal consequences [31] [24].

5.2.3 Imbalance of Information. Another major disadvantage is the lack of collaboration between raw material suppliers, manufacturers, warehouses, wholesalers, and retailers. "The problem is that the different decision-makers do not have access to the same information regarding the state of the entire supply chain network, and in addition they usually operate under different objective functions" [48]. In this decentralized method, manufacturers have a difficult time forecasting demand. In addition, an imbalance of information between supply chain nodes increases cost and stock-outs. However, Nematollahi, Hosseini-Motlagh, and Heydari [35] found that collaborative decision making through information sharing can increase economic benefits for the entire supply chain while also increasing drug fill rate.

5.2.4 Manufacturing Strategy. The mixture of manufacturers 'push' strategy and retailers 'pull' strategy, results in high safety stocks. At any given point, there is usually 4 to 24 weeks of finished goods that has yet to be delivered to patients [49].

5.2.5 Large Network. Medications pass through several nodes before they are delivered to the market. Safety and security issues face organization conflicts as the capital cost to prevent theft and mismanagement is not equally spread across the supply chain. The number of nodes also increases the likelihood for counterfeits to enter the market. Between each node, medications are shipped and handled between multiple parties and often times across national and state borders [49].

5.2.6 Government Regulation. The government heavily regulates pharmaceutical supply chains to ensure a safe and steady supply of medications. The Drug Quality and Security Act [1] signed by President Barack Obama in 2013 introduced new regulations for the manufacturing and the distribution of pharmaceutical products. The policy mandates the creation of systems to trace lot-level transactions and systems to verify product legitimacy. In addition, any company within the supply chain must obtain federal licensure and authenticate the licensure of their trading partners. These required changes place immense financial pressure on pharmaceutical companies, drug distributors, and prescribers to develop sustainable supply chain solutions. The 2023 deadline gives pharmaceutical companies time to test and implement the most sustainable and practical solution [16].

5.2.7 Counterfeits. High inventory levels increase supply chain cost, the potential for theft, and the introduction of counterfeits. It is estimated that 10% of the worldwide pharmaceuticals are counterfeit and approaching 25% in developing countries [30]. Pharmaceutical companies lose an estimated $200 billion annually due to counterfeit drugs [9].

6 BLOCKCHAIN'S POTENTIAL TO MITIGATE THE OPIOID CRISIS

6.1 Moving Opioid Distribution onto the Blockchain

Blockchain can mitigate the opioid crisis through more secure opioid distribution. The 2013 federal passing of The Drug Quality and Security Act [1] provides pharmaceutical supply chain organizations with the necessary regulatory incentives to quickly move onto the blockchain.

The first step to moving opioid distribution onto the blockchain rests in the initial infrastructure investment plan for development and maintenance. The next step is to establish the policies and security clearances of each organization [6]. Once these critical questions are answered, an opioid distribution blockchain would be similar to blockchains in other industries. Each blockchain would start with the genesis node created by the raw material supplier. From there on, each additional downstream node would timestamp an additional hash. When the opioid eventually reaches the patient, the block would contain information on all involved supply chain nodes with timestamps and distribution information including prescribing physician and pharmacist.

The Hyperledger design principles [8] [53], Tapscott and Tapscott's seven design principles for blockcahin [51] and BlockSci [29] analysis protocols should be included in the design of the blockchain.

6.2 Benefits

6.2.1 Cost Savings. As a proactive cost saving maneuver, drug makers and retailers can move onto the supply chain to prevent future litigation [36]. In addition, blockchain automation saves time and operating costs [53].

6.2.2 Reducing Lead Times. Collaborative record-sharing is the foundation and ultimate strength of blockchain technology. Nematollahi, Hosseini-Motlagh, and Heydari [35] show that collaborative

record-sharing among pharmaceutical nodes increases both the social and economic effectiveness of the supply chain. The economic benefits realized through the reduction of the total supply chain inventory levels also decreases lead times.

6.2.3 Collaborative Information Sharing. Blockchain technology has the potential to reduce the opioid epidemic through transparent and decentralized record keeping. In particular, blockchain has the potential to identify prescription drug fraud. Currently without blockchain, opioid addicts can take advantage of the incomplete feedback between physicians and pharmacists by "doctor shopping", modifying, and duplicating prescriptions [14]. With pharmaceutical records on the blockchain, this type of activity is easily identifiable.

Blockchain can reduce illegal opioid prescribing and distribution. In the current centralized record keeping system, the U.S. Drug Enforcement Administration (DEA) relies the Controlled Substances Act of 1970, that requires drug companies to report unusually large or otherwise suspicious orders [22]. Drug makers on the other hand claim their responsibility to report is too vague. As a result, identifying "pill mills" is unnecessarily difficult and time consuming. The DEA's pharmaceutical unit has 600 investigators [22]. With blockchain, record keeping is standardized and accessible to all parties with the correct cryptographic keys.

6.2.4 Post-Sale Opioid Collection. Blockchain technology can also increase the usefulness of post-sale opioid collection. Current medication packaging lacks 2D DataMatrix barcodes making it nearly impossible to identify historical information such as who is returning their medication, who prescribed and sold the medication, and when the medication was prescribed and returned [55]. Blockchain can trace this information leading to better post-sale analysis. In turn, this information can be studied to improve prescribing methodology.

6.2.5 Counterfeit Detection. Blockchain can reduce the high prevalence of illicit counterfeit drugs. Blocks are immutable, that is once a block is created it cannot be deleted or erased [14]. In addition, each batch of product can be traced back to its origin. This means that each batch will have a block of code associated with it. If a batch does not have its presence in the blockchain, then it can be deemed as a counterfeit [50]. Furthermore, blocks with abnormal distribution patterns can be flagged and removed from the supply chain. Creating illicit blocks is easily identifiable as all new blocks must be approved by all parties on the blockchain.

6.2.6 Data Analysis. Academic institutions and researchers should have access to superkeys to analyze trend analysis to provide predictions and usage patterns over various locations and times of the year [34]. Data analysis can provide both a descriptive and predictive overview of the opioid supply chain.

6.3 Outlook

Moving forward, blockchain must overcome multiple adoption risks. Blockchain monopolization in the pharmaceutical supply chain will reduce the effectiveness, safety, and security of the system. Future governmental regulations can prevent mergers and acquisitions in this industry. In addition, future quantum computing power may be strong enough to break cryptographic keys. Investing in security is necessary for future blockchain success. Lost keys will result in irretrievable data; thus, developing a system to overcome this issue is critical. Lastly, blockchain technology is only as good as its users. Encouraging accurate and timely data entry will ultimately define the usability of blockchain in prescription drug distribution [14].

Nonetheless, blockchain can play a role in mitigating the deadly opioid epidemic by providing cost savings, reducing lead times, facilitating information sharing, facilitating post-sale opioid collection, detecting counterfeit, and providing rich data for analysis.

7 CONCLUSION

Although still in its infancy, blockchain has the potential to be just as transformative as TCP/IP. Early and potential applications in healthcare and supply chain suggest that blockchain is indeed moving along the path of technology adoption. Because blockchain is a low-cost solution for supply chain management and provides security and transparency, it can be used for digital data and communication to overall the distribution of controlled substances such as opioids but this model has yet to be tested. But the computation and infrastructure cost for the entire model is low and should be tested to develop a proof of concept system that leverages blockchain to more securely distribute prescription opioids. A prototype model of blockchain can be developed which emulates the current structure of a pharmaceutical supply chain. Such a model can be vital to test out the flaws of blockchain and how to accurately tailor it to the specific use case.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 113th Congress. 2013. H.R.3204 - Drug Quality and Security Act. (Nov. 2013). https://www.congress.gov/bill/113th-congress/house-bill/3204 Sponsor Rep. Fred Upton.

[2] Howard G. Birnbaum, Alan G. White, Matt Schiller, Tracy Waldman, Jody M. Cleveland, and Carl L. Roland. 2011. Societal Costs of Prescription Opioid Abuse, Dependence, and Misuse in the United States. *Pain Medicine* 12, 4 (2011), 657–667. https://doi.org/10.1111/j.1526-4637.2011.01075.x

[3] Paul Brody. 2017. How Blockchain Revolutionizes Supply Chain Management. (Aug. 2017). http://www.digitalistmag.com/finance/2017/08/23/how-the-blockchain-revolutionizes-supply-chain-management-05306209

[4] Scott Burris, Joanna Norland, and Brian R Edlin. 2001. Legal aspects of providing naloxone to heroin users in the United States. *International Journal of Drug Policy* 12, 3 (2001), 237 – 248. http://www.sciencedirect.com/science/article/pii/S0955395901000809

[5] Shamard Charles. 2017. CVS to Limit Opioid Prescriptions to 7-Day Supply. (Sept. 2017). https://www.nbcnews.com/storyline/americas-heroin-epidemic/cvs-limit-opioid-prescriptions-7-day-supply-n803486

[6] K. Christidis and M. Devetsikiotis. 2016. Blockchains and Smart Contracts for the Internet of Things. *IEEE Access* 4 (06 2016), 2292–2303. https://doi.org/10.1109/ACCESS.2016.2566339

[7] Theodore Cicero, Matthew Ellis, Hilary L Surratt, and Steven Kurtz. 2014. The Changing Face of Heroin Use in the United States A Retrospective Analysis of the Past 50 Years. *JAMA psychiatry* 71 (05 2014), E1–E6.

[8] Sharon Cocco and Gari Singh. 2017. Top 6 technical advantages of Hyperledger Fabric for blockchain networks. (Aug. 2017). https://www.ibm.com/developerworks/cloud/library/cl-top-technical-advantages-of-hyperledger-fabric-for-blockchain-networks/index.html

[9] Reenita Das. 2017. *Does Blockchain Have A Place In Healthcare?* Technical Report. Forbes. https://www.forbes.com/sites/reenitadas/2017/05/08/does-blockchain-have-a-place-in-healthcare/#5ebcaa6d1c31

[10] Corey Davis, Traci Green, and Leo Beletsky. 2017. Action, Not Rhetoric, Needed to Reverse the Opioid Overdose Epidemic. *Journal of Law, Medicine*

& Ethics 45 (2017), 20 – 23. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=aph&AN=122737813&site=ehost-live&scope=site

[11] JULIE HIRSCHFELD DAVIS. 2017. Trump Declares Opioid Crisis a fiHealth Emergencyfi but Requests No Funds. (oct 2017). https://www.washingtonpost.com/investigations/cherokee-nation-sues-drug-firms-retailers-for-flooding-communities-with-opioids/2017/04/20/03d04a74-2519-11e7-b503-9d616bd5a305_story.html?utm_term=.ee0423b994ba

[12] The Economist. 2014. *The 272 billion dollar swindle.* Technical Report. The Economist, https://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle.

[13] Ariel Ekblaw, Asaf Azaria, Thiago Vieira, and Andrew Lippman. 2016. *MedRec: Medical Data Management on the Blockchain.* Technical Report. pubpub.org.

[14] Mark A. Engelhardt. 2017. Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector. *Technology Innovation Management Review* 7 (10/2017 2017), 22–34. https://doi.org/10.22215/timreview/1111

[15] Curtis Florence, Chao Zhou, Feijun Luo, and likang xu. 2016. The Economic Burden of Prescription Opioid Overdose, Abuse, and Dependence in the United States, 2013. *Medical Care* 54 (01 2016), 901–906.

[16] U.S. Food and Drug Administration. 2014. Title II of the Drug Quality and Security Act. (Dec. 2014). https://www.fda.gov/Drugs/DrugSafety/DrugIntegrityandSupplyChainSecurity/DrugSupplyChainSecurityAct/ucm427033.htm

[17] Kaiser Family Foundation. 2015. Opioid Overdose Deaths by Race/Ethnicity. (july 2015). https://www.kff.org/other/state-indicator/opioid-overdose-deaths-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[18] Gideon Greenspan. 2016. Four genuine blockchain use cases. (May 2016). https://www.multichain.com/blog/2016/05/four-genuine-blockchain-use-cases

[19] Claire Hansen. 2017. CVS to Limit Opioid Prescriptions. (Sept. 2017). https://www.usnews.com/news/national-news/articles/2017-09-22/cvs-to-enforce-new-limits-on-opioid-prescriptions

[20] Kathryn Hawk, Federico E Vaca, and Gail D'Onofrio. 2015. Reducing Fatal Opioid Overdose: Prevention, Treatment and Harm Reduction Strategies. *The Yale journal of biology and medicine* 88 (09 2015), 235–245.

[21] Scott Higham and Lenny Bernstein. 2017. Cherokee Nation sues drug firms, retailers for flooding communities with opioids. (april 2017). https://www.washingtonpost.com/investigations/cherokee-nation-sues-drug-firms-retailers-for-flooding-communities-with-opioids/2017/04/20/03d04a74-2519-11e7-b503-9d616bd5a305_story.html?utm_term=.ee0423b994ba

[22] Scott Higham and Lenny Bernstein. 2017. THE DRUG INDUSTRYfiS TRIUMPH OVER THE DEA. (Oct. 2017). https://www.washingtonpost.com/graphics/2017/investigations/dea-drug-industry-congress/?utm_term=.86b20fd58ff4

[23] Hoovers. 2017. Drug Wholesalers. (2017). http://subscriber.hoovers.com/H/industry360/overview.html?industryId=1493

[24] Mei-na Hua, Hua-jun Tang, and Zi-lin Wu. 2016. Analysis of a pharmaceutical reverse supply chain based on unwanted medications categories in household. In *Industrial Engineering and Engineering Management (IEEM), 2016 IEEE International Conference on*. IEEE, IEEE, Bali, Indonesia, 1493–1497.

[25] Marco Iansiti and Karim Lakhani. 2017. *The Truth About Blockchain.* Technical Report. Harvard Business Review.

[26] Marco Iansiti and Karim R. Lakhani. 2017. The Truth About Blockchain. (feb 2017). https://hbr.org/2017/01/the-truth-about-blockchain

[27] RFID Journal. 2017. How much does an RFID tag cost today? (Aug. 2017). http://www.rfidjournal.com/faq/show?85

[28] Kost De Serves N. Chilton B Kakavand, H. 2017. The Blockchain Revolution: An Analysis Of Regulation And Technology Related To Distributed Ledger Technologies. (april 2017). http://www.fintechconnectlive.com/wpcontent/uploads/2016/11/Luther-Systems-DLA-Piper-Article-onBlockchain-Regulation-and-Technology-SK.pdf

[29] Harry A. Kalodner, Steven Goldfeder, Alishah Chator, Malte Moser, and Arvind Narayanan. 2017. BlockSci: Design and applications of a blockchain analysis platform. *CoRR* abs/1709.02489 (oct 2017), 1–14. http://arxiv.org/abs/1709.02489

[30] Theodore Kelesidis, Iosif Kelesidis, Petros I. Rafailidis, and Matthew E. Falagas. 2007. Counterfeit or substandard antimicrobial drugs: a review of the scientific evidence. *Journal of Antimicrobial Chemotherapy* 60, 2 (2007), 214–236. https://doi.org/10.1093/jac/dkm109 arXiv:/oup/backfile/content_public/journal/jac/60/2/10.1093_jac_dkm109/1/dkm109.pdf

[31] Peter Kelle, John Woosley, and Helmut Schneider. 2012. Pharmaceutical supply chain specifics and inventory solutions for a hospital case. *Operations Research for Health Care* 1, 2 (2012), 54 – 63. https://doi.org/10.1016/j.orhc.2012.07.001

[32] Jeanne Lenzer. 2014. US could recycle 10 million unused prescription drugs a year. *BMJ* 349 (2014), g7677. https://doi.org/10.1136/bmj.g7677

[33] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big data: The next frontier for innovation, competition, and productivity.* Technical Report. McKinsey Global Institute, https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

[34] Steven McKie. 2015. The Blockchain Meets Big Data and Real-time Analysis. (June 2015). https://bitcoinmagazine.com/articles/blockchain-meets-big-data-realtime-analysis-1435183048

[35] Mohammadreza Nematollahi, Seyyed-Mahdi Hosseini-Motlagh, and Jafar Heydari. 2017. Economic and social collaborative decision-making on visit interval and service level in a two-echelon pharmaceutical supply chain. *Journal of Cleaner Production* 142, Part 4 (2017), 3956 – 3969. https://doi.org/10.1016/j.jclepro.2016.10.062

[36] Yuki Noguchi. 2017. 41 States To Investigate Pharmaceutical Companies Over Opioids. (Sept. 2017). https://www.npr.org/sections/thetwo-way/2017/09/19/552135830/41-states-to-investigate-pharmaceutical-companies-over-opioids

[37] Commission on Combating Drug Addiction and the Opioid Crisis. 2016. Commission Interim Report. (june 2016). https://www.whitehouse.gov/sites/whitehouse.gov/files/ondcp/commission-interim-report.pdf

[38] National Institue on Drug Abuse. 2017. Opioid Crisis. (june 2017). https://www.drugabuse.gov/drugs-abuse/opioids/opioid-crisis#one

[39] Bryce Pardo. 2017. Do more robust prescription drug monitoring programs reduce prescription opioid overdose? *Addiction* 112, 10 (2017), 1773–1783. https://doi.org/10.1111/add.13741 ADD-16-0812.R1.

[40] Stephen W. Patrick, Carrie E. Fry, Timothy F. Jones, and Melinda B. Buntin. 2016. Implementation Of Prescription Drug Monitoring Programs Associated With Reductions In Opioid-Related Death Rates. *Health Affairs* 35, 7 (2016), 1324–1332. https://doi.org/10.1377/hlthaff.2015.1496

[41] Tom Price. 2017. Strategy for Fighting Opioid Crisis. (april 2017). https://www.hhs.gov/about/leadership/secretary/speeches/2017-speeches/secretary-price-announces-hhs-strategy-for-fighting-opioid-crisis/index.html Tom Price's remarks at the National Rx Drug Abuse and Heroin Summit.

[42] Kefa Rabah. 2017. Overview of Blockchain as the Engine of the 4th Industrial Revolution. *Mara Research Journal of Business & Management-ISSN: 2519-1381* 1, 1 (2017), 125–135.

[43] Nate Raymond. 2017. S. Carolina sues OxyContin maker Purdue over deceptive marketing. (Aug. 2017). "https://www.reuters.com/article/south-carolina-purduepharma/s-carolina-sues-oxycontin-maker-purdue-over-deceptive-marketing-idUSL2N1L10S3"

[44] Shari Rudavsky. 2017. Indiana has an opioid crisis. See what the state's leading university is doing to help. (Oct. 2017). https://www.indystar.com/story/news/2017/10/10/state-has-opioid-crisis-see-what-its-leading-university-pledges-50-million-address-opioid-crisi/747151001/

[45] David F Scholl L Rudd RA, Seth P. 2016. Increases in Drug and Opioid Involved Overdose Deaths United States. *MMWR Morb Mortal Wklys* Rep 2016, 65:1445fi?!1452 (May 2016). http://dx.doi.org/10.15585/mmwr.mm655051e1

[46] Lainie Rutkow, Lydia Turner, Eleanor Lucas, Catherine Hwang, and G. Caleb Alexander. 2015. Most Primary Care Physicians Are Aware Of Prescription Drug Monitoring Programs, But Many Find The Data Difficult To Access. *Health Affairs* 34, 3 (2015), 484–492. https://doi.org/10.1377/hlthaff.2014.1085

[47] Krystsina Sadouskaya. 2017. *Adoption of Blockchain Technology in Supply Chain and Logistics.* Master's thesis. Mikkeli University of Applied Sciences.

[48] Nihar Sahay and Marianthi Ierapetritou. 2013. Centralized vs. Decentralized Supply Chain Management Optimization. (11 2013). https://aiche.confex.com/aiche/2013/webprogram/Paper319958.html

[49] Nilay Shah. 2004. Pharmaceutical supply chains: key issues and strategies for optimisation. *Computers & Chemical Engineering* 28, 6 (2004), 929 – 941. https://doi.org/10.1016/j.compchemeng.2003.09.022 FOCAPO 2003 Special issue.

[50] Axel Domeyer Steve Cheng, Matthias Daub and Martin Lundqvist. 2017. Using blockchain to improve data management in the public sector. (Feb. 2017). https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/using-blockchain-to-improve-data-management-in-the-public-sector

[51] Don Tapscott and Alex Tapscott. 2016. *Blockchain Revolution: How the Technology behind Bitcoin is changing Money, Business, and the World.* Penguin Random House LLC, 375 Hudson St, New York, New York 10014.

[52] Douglas Taylor. 2014. RFID in the Pharmaceutical Industry: Addressing Counterfeits with Technology. *Journal of Medical Systems* 38, 11 (12 Oct 2014), 141. https://doi.org/10.1007/s10916-014-0141-y

[53] TheLinuxFoundationProject. 2017. Revolutionizing the Supply Chain. (2017). https://www.hyperledger.org/projects/sawtooth/seafood-case-study

[54] Kevin Vowles, Mindy Mcentee, Peter Siyahhan Julnes, Tessa Frohe, John Ney, and David N van der Goes. 2015. Rates of opioid misuse, abuse, and addiction in chronic pain. *Pain* 156, 4 (04 2015), 569–576.

[55] Dan Walles. 2017. Track and trace is on the way. Is your drug supply chain ready? (June 2017). https://medcitynews.com/2017/06/track-and-trace-are-you-ready/

[56] David C. Wyld. 2008. Genuine medicine?: Why safeguarding the pharmaceutical supply chain from counterfeit drugs with RFID is vital for protecting public

health and the health of the pharmaceutical industry. *Competitiveness Review* 18, 3 (2008), 206–216. https://doi.org/10.1108/10595420810905984

[57] Ben Yuan, Wendy Lin, and Colin McDonnell. 2016. *Blockchains and electronic health records.* Technical Report. MIT.

Big Data and Sentiment Analysis

Syam Sundar Herle
Indiana University
711 N Park Ave
Bloomington, Indiana 47408
syampara@iu.edu

ABSTRACT

Opinion mining is an art of extracting opinions of author, speaker or writer from their written text or spoken words, opinion mining is done by employing Natural Language Processing ,text analysis, Artificial Intelligence. This paper shows where the Big Data meets Sentiment Analysis and the issues related in using Big Data in Sentiment Analysis. How the consumer needs and product evaluation is done while using Big Data in opinion mining.

KEYWORDS

Big Data, i523, hid219, Opinion Mining, Natural Language Processing, Artificial Intelligence, Sentiment Analysis

1 INTRODUCTION

Usually the sentiment analysis can be defined as classifying a piece of written text into positive,negative or neutral state. This is also known as opinion analysis as it derives the opinion of the author. Said that, now a days, user's and consumers write reviews, opinions on online for a product or service given to them. The growth of web and internet has enabled user or consumers to communicate among each other and write up more reviews about products and has made the web more subjective and opinionated. As a result, huge amount of data in the form of texts are available from online which are rich in information, which are users review and opinion about products and service delivered to them. Which in turn can be used by the product owners to meet the consumer needs and demands, and provide quality products and identify the issues in products and much more. Unbiased and independent reviews about a product are the main decision making entity usually used by peoples while purchasing a product. Opinion mining is usually employed for the knowledge retrieval from textual contents such as review and comments. Opinion mining results the the textual content in one of the three categories namely positive, negative and neutral. Positive opinion are good in yielding good financial gains and fame and good business for product owners. On the other hand negative reviews helps in identifying the issues of the products, having said that some people usually write fake reviews to boost their financial gain or discredit their competitive product. These type of reviews are known as opinion spammers.

2 SENTIMENT ANALYSIS

As defined earlier, sentiment analysis or opinion mining is the study of peoples opinion, attitude and emotion towards a product or any other entity. Usually the sentiment analysis is done on the textual contents like reviews of product, twitter tweets and comments in public blog about any entities. The overview of the sentiment analysis can be seen the Figure 1.

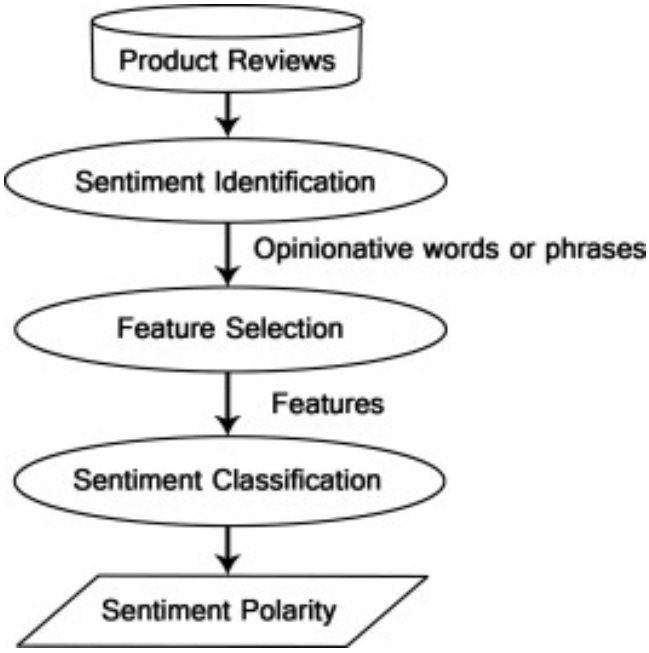


Figure 1: Sentiment Analysis on product review [6]

Overall, sentiment analysis is considered to be a classification process, which is done by three different types, the first one is being document sentiment analysis where the texts of documents are used to classify the opinion of the whole documents, the other type is sentence sentiment analysis where the texts of the sentence is used to classify the opinion of the sentence and third one is being aspect-level sentiment analysis where sentence or document are classified with respect to specific aspects of entity (product or service).

2.1 Data

In the first step of Sentiment Analysis, data are collected from online, the preferred data are reviews of product as these data are textual contents and unstructured and rich in information. These data are important in the sentiment analysis as the business owner can make use of the analysis of these data (review of product) to learn about users opinion about the product. The social media network are one of the good and important source of these types of data as users and consumers interact and post review of the products from their own accounts.

The second important step is identifying the words or phrases which are useful in the forming the opinion in the sentence or

document based on the type of sentiment analysis we intend to do as told in earlier paragraph. The following sub-section describes the next each steps in sentiment analysis in detail.

2.2 Feature Selection

The very important step in sentiment analysis is extracting features needed for the analysis, some of the feature selection used are,**Bag of Words and Frequency** in which a vector of binary is used as features, the binary vector comprises of 1 and 0 based on the presence of words in the document or sentence and forming a bag of binary vector for all words. The other weight used in these type of feature is the frequency of the words in the sentence or document.

The next type of features used in sentiment analysis is **Part Of Speech** where the part of speech tagger is used to create features, in this method grammatical context tag of the words in sentence or document is used, these types features helps to find the emotion of the author from his or her texts.

The other features employed are **Opinion words and Negations** where features are built based upon the words which determines the words as bad or good and negation features gives overall appearance of the negative words in sentence or document.

2.2.1 Feature Selection Methods. Here we can see some of the methods employed to extract and select features to perform sentiment analysis on the sentence or document. Usually feature selection can be done by two types of methods, **Lexicon-based** [8] method and **Statistical-based** [10] method. The first method needs human annotation like bag of words (BOW) [8] but the second method is usually is fully automated. Some of the statistical methods used in feature selection are given as follows,

Chi-Square $\tilde{\chi}^2$. : In chi-square test, the correlation between the term and categories is checked. According to [6] Consider n documents in any collection, and $p_i(w)$ be the conditional probability of class i for document which contains the word w. and P_i be the global fraction of document containing the class i, and $F(w)$ be the global fraction of document containing the word w. So the chi-square can be calculated as given in [6],

$$\tilde{\chi}^2 = \frac{n.F(w)^2.(p_i(w) - P_i)^2}{F(w).(1 - F(w)).P_i.(1 - P_i)}$$

The chi-square gives weights value for the words in the document and these value can be used as features for the sentiment analysis.

Information Gain. : In this method features are created based on the ranks of Information Gain entropy grouped in descending order. "The information gain usually measures the amount of information in bits about the class prediction, it also measures the expected reduction in entropy" [2].

Point-Wise Mutual Information. : This model provides mutual information between feature and classes, and this is derived from the information theory. According to [6] the point-wise mutual information, $M_i(w)$ is the information between the word w and class i. So, we in laymen terms , the PMI [6] is the ratio between expected co-occurrence of class i and w which is given by $F(w).p_i(w)$ and the true occurrence which is given by $F(w).P_i$, so the PMI is defined as,

$$M_i(w) = log\left(\frac{F(w).P_i}{F(w).p_i(w)}\right)$$

The word w is positive correlated to class i if the value of $M_i(w)$ is greater than 0. Comparing with chi-square, PMI is not normalized value and hence for most of the feature selection chi-square is used.

2.3 Classification Technique

The classification technique employed for the sentiment analysis can be done by three approaches, Machine Learning approach, Lexicon based approach and use of both ML and Lexicon approach. In this sub-section we will walk through some of the techniques followed in all the three approaches.

2.3.1 Machine Learning Approach. Typical machine learning algorithms or predictive models are used in this approach, typically the ML approach can be divided into two models, Supervised learning and Unsupervised learning.

Supervised Learning. : In supervised learning, a model is trained with the help of the labeled training data set and evaluated on the unseen data set which are known as test data set. After evaluating the test data classification accuracy are calculated to know how good the trained model is in classifying. Some of the supervised learning model used in sentiment analysis for classification techniques are,

- Naive Bayes classifier : The Naive-Bayes is one of the probabilistic classifier, which works based on probabilities value to determine class. According to [6] the Naive-Bayes model predicts the posterior probability of a class to which the document belongs to, based on distribution of words in the document. For a given features it calculates the probability of the label assuming all the features $(f_i - f_n)$ are independent to each other, the Naive-Bayes follows the following equation to classify the class to which the document belongs to given the features, the features are generated by the Bag Of Words model.

$$P(label/features) = \frac{P(label) * P(f_1/label) * . * P(f_n/label)}{P(features)}$$

  As the above equation assumes the naive assumption of independence and uses Bayes theorem, it is known as the Naive-Bayes model.
- Support Vector Machine : Support Vector Machine model works based on the linear separation of the features which separates features in the search space to separate them based upon their classes, SVM is one of the linear classifier used in supervised learning ML approach. In SVM, the features are transformed to a higher dimensional state space and hyperplanes are used to separate the features, hyperplanes are determined by support vectors, which are features closer to the decision boundary of the separation of the features. For n class of class the model needs $n - 1$ support vectors. "SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane" [6]. In the case opinion mining, the classification are two positive

and negative and hence one support vector is needed to do sentiment analysis.

- Neural Network : Neural Network (NN) consists of many basic unit which are known as neurons, the word frequencies in a document X_i is given as input to the neuron, which has certain weights A to compute probabilities values P_i which is a product of $X_i * A$, the probability values acts as a linear function $f(.)$. So the linear function is given by,

$$f(.) = X_i * A$$

For the classification problem the linear function predicts the class label for the input X_i. The more the layer of neuron the better the output prediction will be, hence Multilayer Neural Network is employed in the classification problem if the number of class is more than 2.

Unsupervised Learning. In the case of unsupervised learning, we will create a model to classify the classes of the data, the model is made to learn from the unlabelled data. In the case of sentiment analysis, analysis is done based on the similarity of the text and clustering them. Some of the unsupervised learning approach used in sentiment analysis are,

- LDA : Latent Dirichlet allocation model is generative statistic model, Xianghua and Guo [11] used LDA model to automatically discover the aspects discussed in Chinese social reviews and also the sentiments expressed in different aspects.
- k-means : k-means employ metric to calculate distance between features created by bag of words (BOW) to cluster the words.

2.3.2 Lexicon Based Approach. : Opinion Lexicons are used in the case of the lexicon based approach. Opinion words or phrases are the words which describe the state or the nature of word. Positive opinion is used for desired state and negative words are used in the case of the undesired state, opinion lexicons are group of pinion words and idioms. The collection of opinion lexicon are done by manually and automated, among which the automated approach is used to collect the opinion lexicons. The automated approach is done by two ways which are discussed in the following sections.

Dictionary Based approach. : In this approach well known corpora like WordNet [7] is employed, initially few opinion lexicon which are specific with context orientation are collected manually and further lexicon are built using corpora.

Corpus Based approach. : In this method syntactic pattern of the seed words (opinion words) are used to find the other opinion words from the corpus. It addresses the problem of manually finding few opinion words in dictionary based approach.

3 BIG DATA

"Big Data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data. While many of the techniques to process and analyze these data types have existed for some time, it has been the massive proliferation of data and the lower cost computing models that have encouraged broader adoption" [9]. Usually Big Data is comprised of four V's as said in [9] which is as follows,

3.1 Volume

Volume refers to the amount of data which are processed and stored, big data usually comprises of the high volume of low-level data which are granular in nature. Real-time data captured from CCTV, internet live feed are so dense and has high volume, big data technologies allows user to convert these low-level high volume to high-level density data.

3.2 Velocity

The speed at which data captured can be defined here. As said in earlier sub-section data acquired from live feed CCTV are real-time in nature so the speed at which data are transmitted to the target storage also plays major role. Big Data requires processing and storage units to keep with the speed of the data being accumulated. Traditional data storage cannot be used to handle Big Data. Big Data technology like Hadoop are required to store data.

3.3 Variety

The real nature or raw nature of the Big Data ranges from structured, semi-structured and unstructured. As the technology have grown, many unstructured data are being captured from social media, internet. Audio and Video files comprises to the most of the unstructured data. Big Data technologies are needed to convert these unstructured to structured data and process it.

3.4 Value

According to [9] Data has intrinsic value, which are quantitative and textual in nature. Appropriate investigation method needed to derive the real value of the data to reveal the knowledge encapsulated in data which ranges from discovering a consumer preference or sentiment, to making a relevant offer by location, or for identifying a piece of equipment that is about to fail. The technological breakthrough is that the cost of data storage and compute has exponentially decreased, thus providing an abundance of data from which statistical sampling and other techniques become relevant, and meaning can be derived.

4 BIG DATA IN SENTIMENT ANALYSIS

In this section we will see some of the Big Data Analytic technology which are used in sentiment analysis. For sentiment analysis, sentiment lexicon or opinion lexicon is needed to classify the text into the different groups which is positive,negative and neutral. One of the method using the opinion lexicon to do sentiment analysis on the text is meta-level feature [1]. In this method, sum on the positive word and sum on the negative word are calculated to classify the text into specific opinion. As this method is manual in nature, a method based on genetic algorithm is proposed in [3] ALGA : Adaptive Lexicon learning using Genetic Algorithm, which creates opinion lexicon dynamically and addresses the manual problem and optimization problems in meta-level feature. But in case if big data is used in sentiment analysis then, ALGA results will be bad, in order to address this, big data analytic technology need to be incorporated [4]. According to [4] the runtime problem in [3] is tackled

by using MapReduce framework of Big Data, by dividing the jobs of calculating scores of positive lexical and negative lexical into independent tasks and run them in parallel on large-scale cluster. In MapReduce framework program, the input data are partitioned into independent set and in Map program the independent set are compiled to produce key and value tuples where key is the word and value is the frequency of the respective word and in the Reduce program the tuples are combined. By doing so, the computational time ALGA [3] is reduced.

Using bigger data usually requires more memory and also more computation time. In order to address this issue, Apache Hadoop [5] was used to store data in efficient manner. Hadoop consists of HDFS file system and MapReduce engine that manages the data storage blocks. HDFS consists of Namenode and Datanode [5] where the Namenode manages all the Datanode and as said earlier MapReduce deals the parallel processing of the data with the help of the master node and slave node. For various classification and clustering problem on the huge data, Apache Mahout [5] is used. For the evaluation process in [5] Apache Mahout splits the data set into training set and test set. The training set is used to train the classifier.

5 SENTIMENT ANALYSIS IN OTHER FIELDS

Now a days, sentiment analysis are applied in many fields, researchers and scientist recently are using sentiment analysis in many fields. Some of the fields where sentiment analysis used are,

5.1 Emotion Detection

Emotion detection is a process of detecting emotion from texts, as sentiment analysis classifies texts into two category positive and negative the sentiment analysis is little different from the emotion detection. Emotion detection is done by either machine learning techniques or lexicon based approach.

5.2 Transfer Learning

Transferring knowledge from one domain to other domain is known as the transfer learning. This method extracts knowledge from one domain and uses the same knowledge in other domain like to search specific text in tweets based on knowledge gained from Wikipedia. Same way sentiment or opinion can be transferred in sentiment analyst classification from one domain to other, it can also be used to build up a bridge between two different domains.

5.3 Subjectivity Classification

In this case sentiment analysis is done to classify text into subjective and objective classes. Text containing a personal view is known to be subjective text but if a text like "Apple released new phone" is a objective text.

6 CONCLUSION

From this paper we learned what is sentiment analysis, how the data are acquired for sentiment analysis and we learned more about feature selection from the data and different types of classification techniques which can be employed in sentiment analysis. Some of the fields using the sentiment analysis was also seen in this paper. This paper also concludes some of the big data analytic used in sentiment analysis and problems that were addressed using the big data in sentiment analysis.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems* 69, Supplement C (2014), 86 – 99. https://doi.org/10.1016/j.knosys.2014.05.016

[2] W lodzis law Duch. 2006. *Filter Methods.* Springer Berlin Heidelberg, Berlin, Heidelberg, 89–117. https://doi.org/10.1007/978-3-540-35488-8_4

[3] Hamidreza Keshavarz and Mohammad Saniee Abadeh. 2017. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems* 122, Supplement C (2017), 1 – 16. https://doi.org/10.1016/j.knosys.2017.01.028

[4] H. Keshavarz, M. S. Abadeh, and M. Almasi. 2017. A new lexicon learning algorithm for sentiment analysis of big data. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, IEEE Journal, 000249–000254. https://doi.org/10.1109/SISY.2017.8080562

[5] M. Kumar and A. Bala. 2016. Analyzing Twitter sentiments through big data. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, IEEE Journal, 2628–2631.

[6] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093 – 1113. https://doi.org/10.1016/j.asej.2014.04.011

[7] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093 – 1113. https://doi.org/10.1016/j.asej.2014.04.011

[8] Phan Trong Ngoc and Myungsik Yoo. 2014. The lexicon-based sentiment analysis for fan page ranking in Facebook. In *The International Conference on Information Networking 2014 (ICOIN2014)*. IEEE, IEEE Journal, 444–448. https://doi.org/10.1109/ICOIN.2014.6799721

[9] Oracle. 2017. Big Data Guide. (nov 2017). http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[10] N. M. Shelke, S. Deshpande, and V. Thakare. 2017. Statistical feature based approach for aspect oriented sentiment analysis. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, IEEE Journal, 376–381. https://doi.org/10.1109/ICICCT.2017.7975223

[11] Fu Xianghua, Liu Guo, Guo Yanyan, and Wang Zhiqiang. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems* 37, Supplement C (2013), 186 – 195. https://doi.org/10.1016/j.knosys.2012.08.003

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."[3]

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-colded 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

2

[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

List of Figures

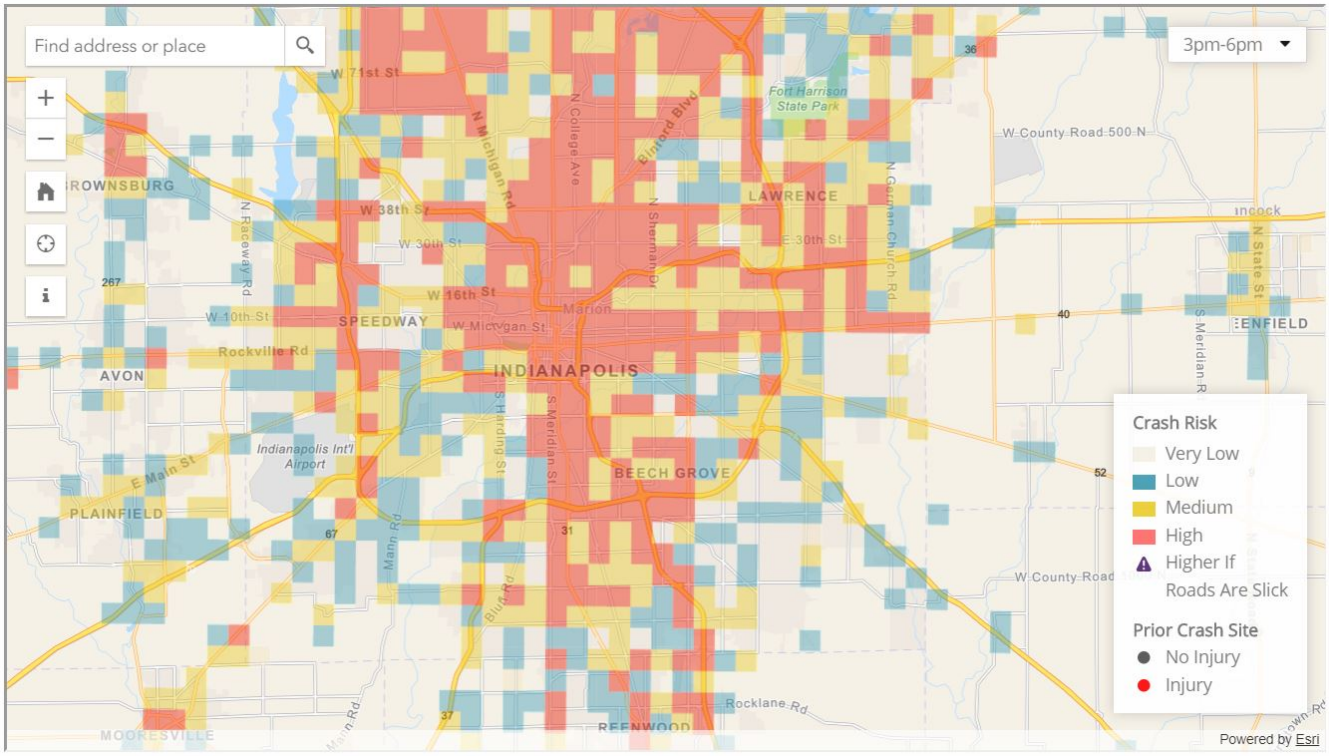


Figure 1: The map shows the probability of a crash through color-coded blocks.[1]

5

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. fiWe are adopting what could be important and impactful practices, but we really donfit know, because we donfit have data to inform, instrument, tune, test, and measure the impact student-level impact of our seemingly endless stream of initiativesfi [5]. How will big data transform education?

3 EDUCATION IS MORE ADAPTIVE

Students with different abilities can learning 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 studentsfi 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 studentsfi learning path; diagnosis of studentsfi needs; reducing studentsfi and institutionfis 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 studentsfi 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. Letfis 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: rpiciples, 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/#201f70a81bcc

[6] Newby Peggy A. Ertmer, Timothy J. 1993. Behaviorism, cognitivism, constructivisim: 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.

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 [7]. 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 [7]. 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 failures of equipments) and customer satisfaction (engaging customers in the process flow by providing them with useful insights about their consumption patterns) [10]. 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" [2].

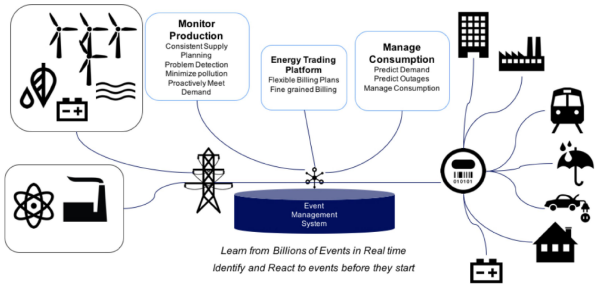


Figure 1: IBM's event-driven Data Management System [7]

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 [21]. 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" [21]. 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

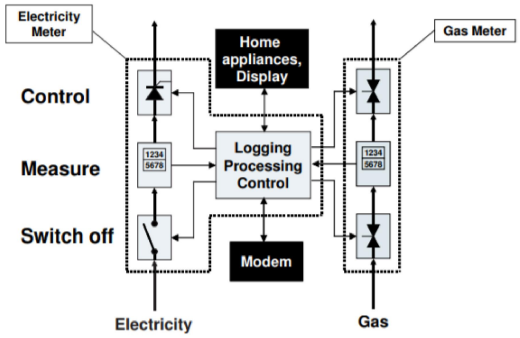


Figure 2: Typical Smart Meter Structure [20]

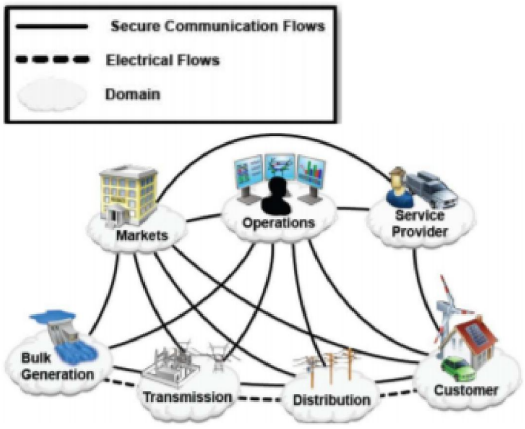


Figure 3: Conceptual model for a Smart Grid [8]

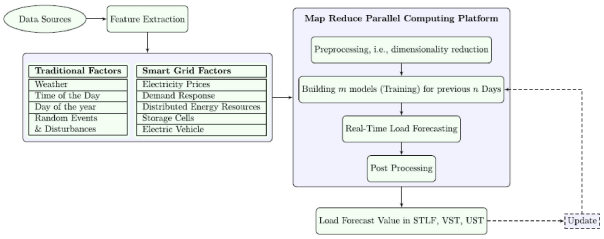


Figure 4: Smart Grid Forecast Model [6]

locally and remotely, remote access of the meter for adjustment of throughput, interconnection among various devices on the premise, ability to read other commodity meters in the vicinity [20]. 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 [20].

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 [15]. 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 [8]. 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 [8]. 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 [8]. 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 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 [6]. 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 [11]. 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 [11].

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 [17]. 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 [17].

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 [16]. 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

Managing its assets and ensuring smooth-running operations are two of the most vital, interdependent tasks for the energy industry. The assets involved in running day-to-day operations are typically quite expensive and result in service interruptions and losses if damaged. Therefore, using prediction and optimization techniques to detect and prevent possible issues in an oil well or power grid or gas pipes is extremely essential, from an asset management point of view as well as to ensure operational efficiency [14]. For example, use of complex algorithms on big data platforms like Hadoop for seismic data analysis can potentially help identify optimal drilling locations and reduce risks in the process [1]. The increase in data being obtained from a variety of sources and the availability of useful machine learning algorithms to make sense of them has opened up opportunities to monitor and detect possible scope of anomalies or leaks and prevent them beforehand. This has also helped in efficient personnel management as it reduces unscheduled visits due to early warnings [18]. Maintenance staff can be better equipped to deal with equipment damages, loads can be shifted to stabilize strained networks and possible outages can be prevented. Predictive maintenance, rather than reactive maintenance, is the need of the hour. An example is Schneider Electric's Avantis PRiSM, a predictive analytics tool, which used prediction based on models built on the asset's operational history during its various phases [18]. In one case, a steam model turbine, which was having sporadic issues over the course of a year was fixed by the maintenance crew by taking corrective measures related to bearing vibrations. However, the issues did not end and the historical data from the turbine was sent to PRiSM for resolution, where it was discovered that thermal expansion issues were the main cause and bearing vibrations just the symptom. Had this predictive maintenance technique been used earlier, initial maintenance would have prevented the machine's issues from becoming chronic and saved millions of dollars on maintenance costs [18].

Another big data analytics methodology has been discussed by Chen, Dokic, Stokes, Goldberg and Kezunovic, on outage prediction [4]. Since weather phenomena are one of the top reasons for outages, the use of a Geographic Information System (GIS) to predict potential outages seems to show great promise. Weather and wind data, vegetation data and electric network data points are consolidated to create a database of all possible influences. This dataset is used along with predictive technologies to create a framework employed to predict risks and plan accordingly [4].

The discussed methodologies provide strong evidence as to how the data generated by the energy sector from its own assets (along with other external factors) can be leveraged to increase operational efficiency. Data from social networks and web browsing data are other potential sources for big data analytics which can be used as complements to existing data sources for inducing useful insights.

3.2 Demand-Side Management and Customer Satisfaction

Demand-Side Management or *Demand-Response Management* is a new approach to load management in the power sector. Generally, supply is adjusted in order to meet demands of the consumers. Demand-side management, however, suggests methods to adjust the demands in order to maintain a balance with the available supply [17]. The most common ways used to implement this are offering financial incentives to reduce demands during peak times or automatically controlling device consumption. The emergence of smart meters and smart grids though, has made this process easier by introducing a two-way network and allowing demand-side management to be applied on the overall electric grid. The use of demand-side management can prove to be useful to the user (in reducing electricity costs) as well as to the supplier (by gaining sufficient time to synchronize the supply and demand, especially if using renewable technologies intermittently) [17]. In a smart grid network, this can be implemented by designing an intelligent system that is able to *read* the grid and provide appropriate responses to changes in load dynamics. Such a system would be able to predict supply and demand trends for the grid and adjust accordingly. Also, the consumers should be willing to allow an intelligent system to manage their devices, by automatically shifting loads when needed, while clearly communicating with the consumer

and asking permissions when required. To develop such a system, advanced big data analytic techniques and simulations are required which can model the grid and all the internal and external factors which affect it [17]. Smart technologies like energy disaggregation can also help by providing the customers with an estimate of their consumption and encouraging them to take preventive measures to prevent overloading on the supplier's end, while reducing their own electricity prices. Factors like the introduction of electric vehicles can also help tremendously in this supply-demand balance strategy (through discharging to satisfy demands and reducing pressure on the grid). The coming together of all these heterogeneous agents (traditional electricity sources, electric vehicles, renewable sources) to maximize energy in the system results in a sort of distributed network often referred to as a *Virtual Power Plant* [17]. The concept of a Virtual Power Plant is essential to demand-side management in its efficiency and ability to increase energy supply while minimizing costs.

Such demand-side management techniques involve high-scale analytics and cloud-computing technologies to deal with the vastness of network information being utilized. Computationally effective algorithms and optimization techniques are crucial to achieve the level of accuracy required. The results however, are invaluable not only to the energy sector but also to the customers through encouragement and implementation of highly efficient energy networks.

3.3 Renewability and Sustainability

The dependence on conventional fuels and sources of energy has increased over time. This has led to an increase in the resultant pollution as well as greater strain on these resources. Renewable energy sources provide alternative and cleaner energy options, thus improving sustainability. However, the issue that renewable energy faces is its intermittent usage. It is still developing as a potential replacement to conventional fuels. Therein lies the question of predicting energy production from these sources. The data generated here is voluminous and varied, therefore making Big Data technologies the best option for analysis purposes. An example is a Vi-POC (Virtual Power Operating Center) designed to provide real-time forecasts of renewable energy generation [3]. The system collects data from various power plants (wind, photovoltaic, biomass or geothermal) along with weather data to predict the generation of energy on an aggregate and individual level. It takes care of the correlation between weather phenomena and internal factors related to the energy production system and employs a semi-supervised adaptive model for forecasting. Mondrian is used for online analytic services, Hive on Hadoop MapReduce is used as a query executor and HBase is used for data storage [3]. The result is an efficient analytics system that provides useful intuition regarding renewable energy production. Deployment and utilization of such technologies can lead to a sustainable blend of renewable and conventional energy sources, thus leading to an energy-efficient framework.

4 WATER MANAGEMENT

Water is a resource imperative to our survival and therefore, one of the most important utility industries. Just as big data analytics has permeated the realm of energy industries, water too is an important customer. The water utility sector has always been quite fragmented, thus the arrival of big data analytics in this field is relatively new. However, there is tremendous scope for improvement and advancement in this sector as well. The concept of smart meters and grids as applied to energy can be used for water resource networks as well. In fact, we do have a lot of data available in the water industry. "Water utilities see data from supervisory control and data acquisition (SCADA) systems, including flow statistics, online monitoring, dissolved oxygen (DO) measurements, and air flows, as well as data from laboratory information management systems (LIMS) and computerized maintenance management systems (CMMS), to name several examples" [19]. Significant use of this data for analytics has also begun. An excellent example can be made of Black and Veatch, an engineering company, which has developed operations optimization tools for the wastewater treatment facility at the city of Lawrence, Kansas. These will be extended further to the water treatment facility as well. The beneficial effects of consolidating all operations data in a single huge database have started becoming evident, especially for visualization purposes [19]. Another shining example is of Fathom, a startup based mainly in the water utility data management area [9]. The CEO of Fathom, Trevor Hill, has led the campaign to revolutionize the water utility sector by setting up a cloud-based platform to offer software-as-a-service to water utilities. Fathom provides meter-reading services on the cloud, automated billing and customer services comprising of leak and meter failure detection. Trevor Hill describes his system as a "smart grid for water utilities" [9]. The rise of smart technologies is gradually, but definitely, reforming the water management sector along with other utilities.

5 CASE STUDIES

Various companies in the energy and utilities sector have used big data analytics to improve their operations and customer management strategies. Each success story reveals how much big data has come along in revamping the entire industry. CenterPoint, a Houston based electric and natural gas utility, has used big data in saving itself a considerable amount of manual labor as well as costs. Physical inspections used to cost the company 88,000 meters every day (75 USD each). This changed with the deployment of smart meters which covered 221 million readings each day and with no traveling costs at all [21]. Oncor Electric Delivery Company LLC, the biggest electricity distribution-transmission company in Texas, collaborated with IBM and installed smart meters for its customers, thus creating an efficient and interactive energy grid and saving its consumers 25 percent and more in electricity bills. Additionally it improved its crisis-response time by almost 40 percent [13]. Vestas, one of the largest wind energy companies in the world, worked with IBM to implement a system that helped in locating optimal turbine placement sites and forecasting wind patterns and energy production. This system produced a 97 percent faster prediction and response time and resulted in 40 percent decrease in energy utilization while increasing computational capacity [12]. Podo, a Spanish utilities company, worked with Cloudera to design a prediction system for consumer energy consumption patterns. This model helped development of micro-targeted campaigns and reduced electricity bills for customers and companies up to 30 percent

[5]. The *big data leap* for all these companies translated into profits and a colossal increase in operational efficiencies.

6 CONCLUSIONS

Implementation of big data technologies is not without its obstacles. Silos mentality and fragmented systems, along with the difficulties in consolidation of highly skilled data scientists and big data based system resources, do pose hindrances on the way to an intelligent and efficient framework. However, we see an increasing number of success stories, people willing to put in the extra effort, to develop intelligence in their systems by breaking the fetters of ignorance. With the increase in technical innovations, it is highly unlikely that the energy and utilities sector will stay behind. The current pace of big data analytics in the industry can assure us that *smart systems* are here to stay.

REFERENCES

[1] Accenture. 2013. *Digitizing Energy - Analytics-Powered Performance.* Accenture.
[2] Stephen Callahan. 2015. Big Data : The Future of Energy and Utilities. (Oct 2015).
[3] Michelangelo Ceci, Nunziato Cassavia, Roberto Corizzo, Pietro Dicosta, Donato Malerba, Gaspare Maria, Elio Masciari, and Camillo Pastura. 2014. Big Data Techniques For Renewable Energy Market. In *22nd Italian Symposium on Advanced Database Systems, SEBD 2014*, Sergio Greco and Antonio Picariello (Eds.), Vol. 1. Italy, 369–377.
[4] Po-Chen Chen, Tatjana Dokic, Nicholas Stokes, Daniel W. Goldberg, and Mladen Kezunovic. 2015. Predicting weather-associated impacts in outage management utilizing the GIS framework. In *2015 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT LATAM)*, Vol. 1. Montevideo, 417–422. https://doi.org/10.1109/ISGT-LA.2015.7381191
[5] Cloudera. 2017. *Podo: Capturing and Growing Market Share with 10x More Accurate Forecasts.* Cloudera.
[6] Panagiotis D. Diamantoulakis, Vasileios M. Kapinas, and George K. Karagiannidis. 2015. Big Data Analytics for Dynamic Energy Management in Smart Grids. *Big Data Research* 2, 3 (Apr 2015), 94–101. https://doi.org/10.1016/j.bdr.2015.03.003
[7] Phil Downey. 2017. Deliver more intelligence to intelligent energy systems. (Aug 2017).
[8] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. 2011. Smart Grid - The New and Improved Power Grid : A Survey. *IEEE Communications Surveys and Tutorials* 14, 4 (Dec 2011), 944–980. https://doi.org/10.1109/SURV.2011.101911.00087
[9] Barbara Grady. 2016. Can Big Data save our water infrastructure? (Feb 2016).
[10] Christopher Guille and Stephan Zech. 2016. How Utilities Are Deploying Data Analytics Now. (Aug 2016).
[11] Wan He and Ying Chai. 2016. An Empirical Study on Energy Disaggregation via Deep Learning. In *Advances in Intelligent Systems Research - 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)*, R.E. Sehiemy, M.B.I. Reaz, and C.J. Lee (Eds.), Vol. 133. Atlantis Press, Netherlands, 338–342.
[12] IBM. 2011. *Vestas - Turning climate into capital with big data.* IBM.
[13] IBM. 2013. *Oncor delivers customer benefits through an integrated smart grid.* IBM.
[14] Infosys. 2017. Energy and Utilities. (2017).
[15] Oracle. 2009. *Smart Metering for Water Utilities.* Oracle.
[16] Oracle. 2013. *Utilities and Big Data: A Seismic Shift is Beginning.* Oracle.
[17] Sarvapali D. Ramchurn, Perukrishnen Vytelingum, Alex Rogers, and Nicholas R. Jennings. 2012. Putting the Smarts Into the Smart Grid: A Grand Challenge for Artificial Intelligence. *Commun. ACM* 55, 4 (Apr 2012), 86–97. https://doi.org/10.1145/2133806.2133825
[18] SchneiderElectric. 2015. *Predictive Asset Analytics at Power Utilities.* Schneider Electric.
[19] Andy Shaw. 2017. Understanding Big Data in the Water Industry. (Mar 2017).
[20] Rob van Gerwen, Saskia Jaarsma, and Rob Wilhite. 2006. Smart Metering. (Jul 2006).
[21] Wharton. 2014. Big Data and Energy : A Clear Synergy. (Sep 2014).

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 device measures to combat it [5]. It it 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].

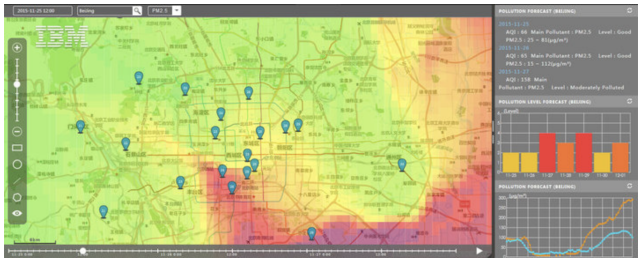


Figure 1: Green Horizons air quality management for Beijing [3]

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 [3].

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].

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]

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 μm PM_{10}, air particles smaller than 2.5 μm $PM_{2.5}$, carbon monoxide CO, and ozone O_3 [6]. Based on the level of AQI, the air quality is classified as shown in Figure 2 [6].

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].

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

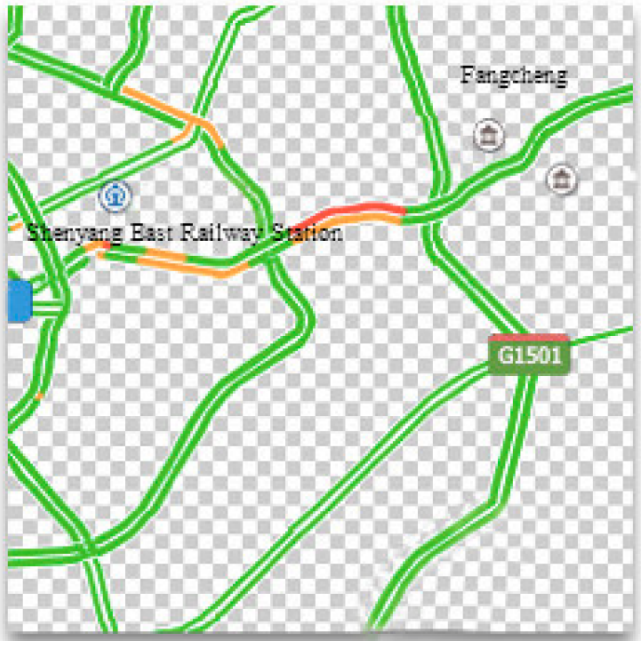


Figure 3: Traffic Congestion[6]

temperature Numeric	humility 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]

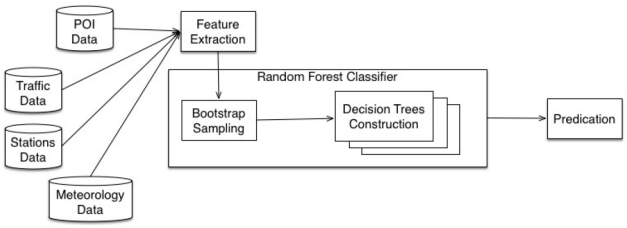


Figure 5: Procedure of RAQ [6]

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 [6].

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 [6].

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-tress 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

Algorithm 1. RAQ

Input:	A dataset S with features: F_{mt}, F_{mh}, F_{mp}, F_{mw}, F_{mv}, F_{ri}, F_{tcs}, F_{ph} 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]

index is determined as the maximum value out of all the individual values [6]. Figure 6 shows the step-by-step RAQ algorithm [6].

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 3.7×10^{-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 [4].

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

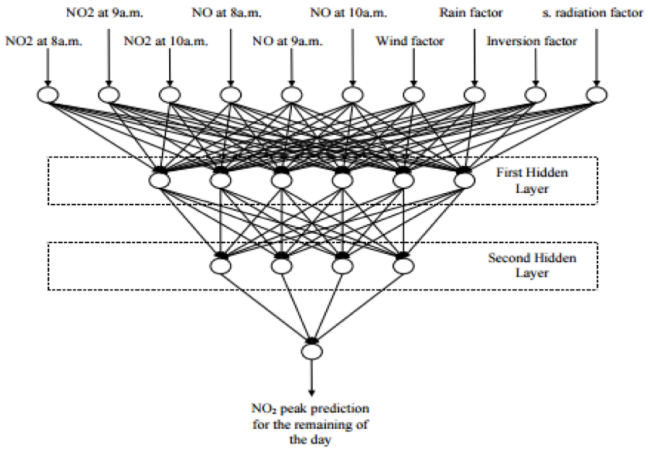


Figure 7: Artificial Neural Network(ANN) Model [4]

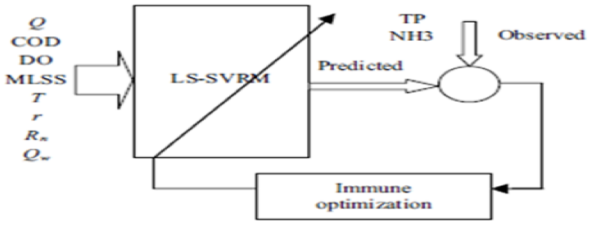


Figure 8: Least squares Support Vector Machine Model [4]

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 consine 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 [4].

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 Oguti 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

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 and a good article to explain the rainwater harvesting

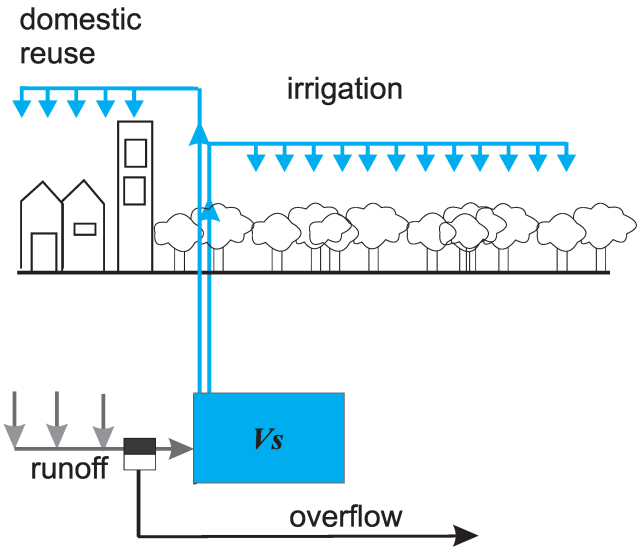


Figure 1: rainwater harvesting figure

process can be found here [4]. The first wash step, in particular, is very important because it removes dust debris etc from the rooftop or else we risk contamination of water. Quite naturally we cannot allow the collected rainwater to overfill tanks in case of this we need to redirect the water to some other outlet.Big data can play a very influential role right from the feasibility analysis to re direction of water.

2.2 Big data and feasibility of installation

Big data can play a huge role in the feasibility estimation. This can be useful for both households and governments, in many countries in some states it is mandatory that each house should have a rainwater harvesting unit. In some cases, these are funded by the government and in some cases it in on the owner to do so. In the case of governments to do an analysis how can they do so, it is a huge task to go to each and every house and track roof dimensions. One easy way of going about this would be to use satellite image data. These images can then be searched for roof features and dimensions accordingly extrapolated and in such a manner the dimensions of many roofs can be obtained and a cost estimate can be obtained. To obtain the data we can use the many datasets

provided by NASA or we can even use a highly zoomed street view from google maps. To extract the features we can use one of the many open CV libraries or use apply complex ML deep learning algorithms . To store this data a simple Hadoop map-reduce can be used. A more detailed study can be viewed in [2].

2.3 Big data and first wash

the importance of first wash was previously stressed upon in the introduction. it is required to wash away dust, debris, dead insects and other such contaminants. The first watch is a manual task it is dependent on the owner to redirect the first wash water elsewhere. Often most people have mistaken it to be the first wash to be the first rain, that is people waste a whole day of rain at times as first wash, or some mistakenly use small drizzles as the first and do not use the first watch properly. Big data can help in the automation of the first wash process combined with IoT. The first step would be to obtain the weather data, this can be achieved either by using the highly consolidated data obtained from the respective government's meteorology departments or the huge datasets provided by the NASA satellite. Once again Hadoop map-reduce allows for reducing a humongous data set into more compact usable structures. From this we can perform a weather data analysis to determine if the rain will be heavy/light and its duration. From this data, we can easily determine when to perform the first watch. Assuming every rainwater harvesting unit has an IoT feature that controls the valves or water redirection one central control center can send signals to a wide area on when to perform the first watch and for how long. This should greatly reduce the amount of rainwater wasted.

2.4 Big data and water tanks

Big data and IoT can once again help in the rainwater harvesting process, there are many times water left in the tanks are not used and the water becomes stagnant with the use of devices the quality of water can be checked and the tanks can be drained. Also, it is very difficult to combine the rainwater tanks with the main water channels of the buildings since the water in the tanks is very very limited. With the help of IoT redistributing this water becomes very easy using data from other parts of the housing analysis can be made water can be redistributed accordingly. Then there is also the matter of making sure the tanks don't overflow which could lead to bursting. Using IoT the water can be tracked in a smart manner and decisions like when to reroute can be done in a smart manner. There is one more important use for big data in tanks, leakages and rusting. As previously mentioned the quality of water can be checked by its ph level using smart devices the next issue comes down to leakages. more often than not most installations are buried under the ground this is done in order to reduce the effects of weather and also so that the installation doesn't take up space. While this leads to a new set of problems the major one is that often leaks cant be detected until its too late. IoT devices can be used to alert a user that water levels are falling down way too rapidly and the user can contact servicemen in time before the next rains.

2.5 Big data and water re routing

Lastly but perhaps most important is the issue of rerouting water once the tanks get full. In the majority of the cases, the rainwater is directly diverted to the water table. While it is normally a good idea to replenish the water table in such a manner due to over exploitation from bore wells. Aside from restoring the water table, it is slowly becoming essential to recharge even the lakes and other sources of freshwater. This is becoming important because with global warming and rainfall becoming more erratic [some places receiving more rainfall than the others and others receiving way lesser] as a result we need to divert some of the harvested rainwater to other lakes/reservoirs. As to how this can be achieved we can use Big data to monitor the water levels and then decide accordingly where to route the rainwater and by how much.

3 TECH IN RAIN WATER HARVESTING

Surprisingly there is not much to write about about very few players exist in the rainwater harvesting market who aim to offer the services of big data. Part of this can be attributed to the fact there is not much data available. For example, Indian government offers highly consolidated annual precipitation data for free while this is useful to perform past estimates it however is really not enough, more detailed minute by minute data of precipitation is required in order for the above mentioned analysis to take place. Even the NASA weather data doesn't give the whole picture. This doesn't mean the data is not there but a premium is required to obtained it. there are a few players who offer smart tank service . However the tech scene is just getting warmed up and awareness of its potential is doing rounds a good article was recently released by NASA on this [1] and a few examples are [3].

4 CONCLUSION

The scope for big data in rainwater harvesting is immense but the major initative lives with the governments, rerouting water into reservoirs is not in the best interests of private contractors but they can be hired to make it so. This can create a huge job market and it will be beneficial for all involved. This also needs to be done sooner rather than later because the rate of population growth exceeds our available sources and if we want the future generations to have any resources we need to embrace technology and start protecting our assets. Also, for private contractors to approach the government with proposals it would be very useful if more relevant data was made available to the public and thus there is a need for investment in the weather department for data.Rainwater harvesting despite being one of the oldest practices of water replenishment is surprisingly behind in terms of technology advancement when we look at the progress made with solar and wind.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ashley Morrow. 2015. Using NASA Data to Show How Raindrops Could Save Rupees. *NASA* 1, 1 (Jan. 2015), 1. https://www.nasa.gov/feature/goddard/using-nasa-data-to-show-how-raindrops-could-save-rupees

[2] Robert O. Ojwang. 2015. Rooftop Rainwater Harvesting for Mombasa: Scenario Development with Image Classification and Water Resources Simulation. *Water 2017* 1, 1 (Jan. 2015), 1. http://www.mdpi.com/2073-4441/9/5/359/htm

[3] UNEP. 2017. rainwater harvesting examples. *unep* 1, 1 (Jan. 2017), 1. http://www.unep.or.jp/ietc/publications/urban/urbanenv-2/9.asp

[4] Wikipedia. 2016. Rain water harvesting. *wikipedia* 1, 1 (Jan. 2016), 1. https://en.wikipedia.org/wiki/Rainwater_harvesting

3

Big Data Applications in Laboratories

Tiffany Fabianac
Indiana University
Bloomington, Indiana 47408, USA
tifabi@iu.edu

ABSTRACT

Ground breaking scientific research and development happen in laboratories all over the world every day. The recent flux of data has revolutionized laboratories across several very different sectors. Many laboratories operate at the capacity that require big data tools. Exploring the current need for big data tools across several industries provides a view of where these tools are currently being applied and how they are benefiting the industry as well as where gaps exist.

KEYWORDS

Big Data, HID313, i523

1 INTRODUCTION

A laboratory is a room or facility designed to conduct experimentation, research, teaching, or manufacturing. Experimentation is the process of performing a defined procedure or test to validate a hypothesis. Most commonly during experimentation, modifications or additions are made to a sample of process to determine the result. Research is the investigation of behavior, material, or process. Research often involves experimentation but does not have to. Teaching within a laboratory introduces students research and experimentation while exploring processes and demonstrating technique. The manufacturing of drugs and medical equipment, the refinement of chemicals such as oil, and food processing are all carried out in laboratories.

There are many different types of laboratories. Analytical laboratories explore the chemical composition of molecules, chemicals, products, and other samples. Biosafety laboratories are designed to offer containment of potentially hazardous chemicals or pathogens. Cleanrooms are designed to protect the elements within the laboratory from airborne particulates. Clinical laboratories perform diagnostic testing and are designed to contain pathological hazards. Production laboratories also require an environment restricting containment and air quality because these laboratories produce very pure and consistent products such as drugs, airplane fuel, or dairy products. There is a vast number of different types of research and development (R&D) laboratories ranging from atomic research labs to laser research labs to mechanical testing lab [7].

Big data is data so numerous that the cost of storing it becomes a burden, it is data that grows exponentially and continuously, and it is data that comes in structured and unstructured forms. Big data provides a source of insight into the elements of cost, time, and process [16]. The push to big data has required the development of software tools that can handle the data load and provide a view into the valuable insights provided.

Where does Big Data meet laboratories? A great majority of laboratories have adopted the digital age with the implementation of Laboratory Information Management Systems (LIMS) and Electronic Laboratory Notebooks (ELN), but many of these laboratories do not produce data on the scale of big data. Only very specialized laboratories are currently producing enormous volumes of structured and unstructured data. Big data tools for laboratories still have number of hurdles to over come such as security, the enormous variability in data diversity, and the steep learning curve but entire industries are coming together to solve these problems in an effort to unlock the power of data [2].

2 CLINICAL LABORATORIES

Clinical laboratories are within the healthcare sector. The specialty is testing the checmical components of body fluid and tissue. Thousands of clinical tests exist and clinical laboratories must be equip and have the ability to run a great number of them quickly [3]. The healthcare industry has been slow to adopt digital solutions which means that many clinical laboratories are still run out of paper notebooks. Data solutions such as Electronic Health Records (EHR), LIMS, ELN, and clinical decision support systems are helping the industry to test samples faster and treat patients smarter.

Big data in the healthcare industry in being applied mainly to electronic health record systems as whole countries find interest in adopting systems that are capable of tracking patient data for the entire population. One particularly interesting use of big data for medical purposes is to analyze electrocardiograph (ECG) data using Hadoop. ECG data is essentially a repetitive time series of a patient's heart activity. Analyzing multiple patient files becomes a daunting task, but using Hadoop allows for ease of storage and the application of time series analysis to identify trends that enable earlier diagnosis of heart disease [18].

The data collected from the many tests run in clinical laboratories has the potential to be combine populations spanning millions, even billions, of patients. A lot of this data will come from ELNs and electronic health record systems which means a lot of unstructured data from free form fields. The collection of this data is allowing for the identification of sub-populations with higher infant mortality, increased risk of cancer, decreased life expectancy, and the like [13]. Laboratory data can then gleam light on what indicators are present to help explain and reduce these health risks. Big Data applications such as Hadoop, Intel Galileo, and cloud plaforms Google Cloud Platform and Amazon Web Services are changing the healthcare industry to combine data in EHR, clinician notes, imaging data, and genomics data [6].

3 PHARMACEUTICAL LABORATORIES

The pharmaceutical sector began to adopt big data system more rapidly then the healthcare industry. Pharmaceutical laboratories specialize in the purification, manufacture, and testing of drugs. Laboratories within this sector are digitized with LIMS and ELN

with a movement towards platfrom solutions that can combine the structured data of LIMS with the unstructured data from an ELN to drive insights. Amazon Webservices and Google Cloud Platform are top of the list within this highly regulated sector because of their options to deploy a validated cloud solution.

A unique aspect of drug development is the creation animal models to assist in validating drug targets. The development and propagation of animal models require genetic data on the scale of big data. High throughput screening is the process of identifying hundreds of genetic or protein markers from a sample. Aside from the storage of results, the analysis of high throughput screening is incomprehensible without the aid of statistical tools [17].

SAP HANA is a cloud based analytic and storage application that has been successfully implemented within the pharmaceutical industry. HANA can be designed to provide accurate sample tracking using Radio Frequency Identification (RFID), detail supply chain, store data on the scale of big data, and perform in depth analysis [4]. Big data solutions are currently in development to apply machine learning algorithms and data mining techniques to the task of drug repositioning. This task involves analyzing collected data toxicology, clinical trial data, published data, drug compatibility data, and more for the purpose to identify new uses for old drugs [19].

4 AEROSPACE LABORATORIES

Not all laboratories are based in biological or chemical sciences. Some labs have telescopes instead of microscopes. Some lab testing starts with a computer model. This is the case in the aerospace industry. An industry that has fully embraced the power of big data applications. Aerospace laboratory testing consists of mechanical analysis, satellite data, physics based models, and advanced computer engineering. In comparison to the healthcare industry, the aerospace industry is on another planet when it comes to big data [5].

Satellite data archives maintain large volumes of observational data. Data and Information Management Systems (DIMS) are designed to handle the data load as well as maintain constant network connection to ensure no part of the real time data feed is lost. The daily data throughput can be from 250 to over 900 GB which means that transferring data requires an even bigger processing engine with some data engines exceeding 10 TB per day transfer rates [9]. 10 TB per day might seem pretty fast until it is considered that NASA hosts a data archive of satellite data that exceeds 500 TB [11]. All this data on its own does not produce much value. The value comes from the analysis of the data.

IBM has developed the Physical Analytics Integrates Data Repository and Services (PAIRS) as a geospatial data repository and analysis engine. The platfrom functions off of IBM's cloud system which stores data on a distributed Hadoop/Hbase system. PAIRS provides the ability to perform time series analysis on satellite and drone images [12]. Apache Spark provides the framework for machine learning algorithms to use and the development of GeoSpark has enabled the process of spatial data as well [11].

5 RESOURCES LABORATORY

The resources industry consists of a lot of mining: coal mining, oil mining, and data mining. Laboratories within the resources sector are devoted to purification and manufacturing tasks. Resources labs measure system efficiency and monitor mechanical functions. Machine learning techniques have been able to identify system failure through auditory and mechanical vibration data [10]. Laboratories within the resources space also adopt LIMS and ELN technologies. The data collected by the oil and gas company Cheveron exceeds 1.5 TB per day [1].

Big data analysis within the resources sector has contributed to optimizing production and saving energy [8]. Oracle, IBM, Hitachi, and Microsoft have all dedicated significant resources to developing big data solutions specific to the resources industry through platform solutions for storage and analysis of data. Syncsort, Pentaho, and Talend have developed analytics tools to run interactive analytics specific to the sector [1]. Cloud technologies present convenience of storage and scaling, but pose the question of security. Companies within the resources industry have shifted towards cloud environments, but prefer private data storage and are even exploring modular IT architectures that support additional storage security [14].

Like the aerospace industry, the resources industry also uses geospatial and geologic data for production. The benefit of big data analysis tools to the industry is the ability to monitor current production as well as identify new patterns and predictions [15].

6 CONCLUSION

Laboratories of all shapes and sizes can use the many big data storage and analysis products to help make scientific decisions and drive innovation throughout the many industries that invest in research and development. Healthcare might be the slowest adopter of big data technologies while the aerospace industry has been using big data storage and analysis tools for decades. As the demand for specialized tools continue to grow within these industries, the products will continue to develop and drive more powerful insights.

Many of the tools currently being used to implement the digitization of laboratories: LIMS, ELN, EHR; are not currently designed with big data in mind. Platform solutions are allowing these systems to grow to the scale of big data. A single hospital or a single clinical trial produces a lot of data, but nothing on the scale of big data. It is when the data from an entire country's population is combined to produce insight that big data tools can really drive decisions. It is when the data from thousands of clinical trials is combined that big data can identify possible targets or alternate uses for drugs. Big data is revolutionizing laboratories in every sector and innovation will only continue to get faster, better, ad smarter.

ACKNOWLEDGMENTS

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

REFERENCES

[1] R. M. Alguliyev, R. M. Aliguliyev, and M. S. Hajirahimova. 2016. Big data integration architectural concepts for oil and gas industry. In *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. 2016 IEEE 10th International Conference on Application of Information and Communication Technologies, Azerbaijan, Baku, 1–5. https://doi.org/10.1109/ICAICT.2016.7991832

[2] C. A. Ardagna, P. Ceravolo, and E. Damiani. 2016. Big data analytics as-a-service: Issues and challenges. In *2016 IEEE International Conference on Big Data (Big Data)*. 2016 IEEE International Conference on Big Data, Washington DC, USA, 3638–3644. https://doi.org/10.1109/BigData.2016.7841029

[3] Standford Health Care. 2014. Laboratory Tests. Website. (12 2014). https://stanfordhealthcare.org/medical-tests/l/lab-tests.html

[4] A. M. Chircu, E. Sultanow, and F. C. Chircu. 2014. Cloud Computing for Big Data Entrepreneurship in the Supply Chain: Using SAP HANA for Pharmaceutical Track-and-Trace Analytics. In *2014 IEEE World Congress on Services*. IEEE World Congress on Services, Anchorage, Alaska, 450–451. https://doi.org/10.1109/SERVICES.2014.84

[5] EVAN DASHEVSKY. 2017. SPACE FOR SALE: How the private space industry will reinvent economics, exploration, and humanity. *PC Magazine* 1, 1 (Aug 2017), 77 – 88. https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com.ezproxy.net.ucf.edu/login.aspx?direct=true&db=aci&AN=124364241&site=eds-live&scope=site

[6] P. Dineshkumar, R.S.Ponmagal, R. SenthilKumar, K. Sujatha, and V.N.Rajavarman. 2016. Big data analytics of IoT based Health care monitoring system. *2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on* 1, 55 (2016), 55. https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com.ezproxy.net.ucf.edu/login.aspx?direct=true&db=edseee&AN=edseee.7894624&site=eds-live&scope=site

[7] Exilab. 2017. Laboratory Types. Website. (01 2017). https://www.exilab.com/en/laboratory-types/

[8] M. Guizhi, Z. Zhanmin, J. Xuefeng, W. Yu, Y. Xizhi, G. Peng, D. Lihong, W. Zanmei, N. Xiaoxia, T. Fujun, Z. Zhaojing, and W. Houbing. 2017. Application of big data analysis in oil production engineering. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*. 2017 IEEE 2nd International Conference on Big Data Analysis, Beijing, China, 447–451. https://doi.org/10.1109/ICBDA.2017.8078859

[9] S. Kiemle, K. Molch, S. Schropp, N. Weiland, and E. Mikusch. 2016. Big Data Management in Earth Observation: The German satellite data archive at the German Aerospace Center. *IEEE Geoscience and Remote Sensing Magazine* 4, 3 (Sept 2016), 51–58. https://doi.org/10.1109/MGRS.2016.2541306

[10] Yaguo Lei, Feng Jia, Jing Lin, Saibo Xing, and Steven X. Ding. 2016. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. *IEEE Transactions on Industrial Electronics* 63, 5 (2016), 3137 – 3147. https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com.ezproxy.net.ucf.edu/login.aspx?direct=true&db=aci&AN=114509151&site=eds-live&scope=site

[11] R. K. Lenka, R. K. Barik, N. Gupta, S. M. Ali, A. Rath, and H. Dubey. 2016. Comparative analysis of SpatialHadoop and GeoSpark for geospatial big data analytics. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. 2016 2nd International Conference on Contemporary Computing and Informatics, Noida, India, 484–488. https://doi.org/10.1109/IC3I.2016.7918013

[12] S. Lu, X. Shao, M. Freitag, L. J. Klein, J. Renwick, F. J. Marianno, C. Albrecht, and H. F. Hamann. 2016. IBM PAIRS curated big data service for accelerated geospatial data analytics and discovery. In *2016 IEEE International Conference on Big Data (Big Data)*. 2016 IEEE International Conference on Big Data, Washington DC, USA, 2672–2675. https://doi.org/10.1109/BigData.2016.7840910

[13] M. Panda, S. M. Ali, and S. K. Panda. 2017. Big data in health care: A mobile based solution. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. 2017 International Conference on Big Data Analytics and Computational Intelligence, Chirala, India, 149–152. https://doi.org/10.1109/ICBDACI.2017.8070826

[14] Robert K. Perrons and Adam Hems. 2013. Cloud computing in the upstream oil & gas industry: A proposed way forward. *Energy Policy* 56, Supplement C (May 2013), 732 – 737. https://doi.org/10.1016/j.enpol.2013.01.016

[15] Robert K. Perrons and Jesse W. Jensen. 2015. Data as an asset: What the oil and gas sector can learn from other industries about fiBig Datafi. *Energy Policy* 81, Supplement C (Jun 2015), 117 – 121. https://doi.org/10.1016/j.enpol.2015.02.020

[16] SAS. 2017. Big Data What it is and why it matters. Website. (08 2017). https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[17] C. Seebode, M. Ort, C. Regenbrecht, and M. Peuker. 2013. BIG DATA infrastructures for pharmaceutical research. In *2013 IEEE International Conference on Big Data*. 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 59–63. https://doi.org/10.1109/BigData.2013.6691759

[18] J. Sivaranjani and A. N. Madheswari. 2017. A novel technique of motif discovery for medical big data using hadoop. In *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*. 2017 Conference on Emerging Devices and Smart Systems, Tiruchengode, India, 214–217. https://doi.org/10.1109/ICEDSS.2017.8073683

[19] Jia Zhang, Candong Li, Yaojin Lin, Youwei Shao, and Shaozi Li. 2017. Computational drug repositioning using collaborative filtering via multi-source fusion. *Expert Systems with Applications* 84, Supplement C (May 2017), 281 – 289. https://doi.org/10.1016/j.eswa.2017.05.004

Concussions and Big Data's Opportunities and Challenges

Jeffry L. Garner
Indiana University
Online Student
jeffgarn@iu.edu

ABSTRACT

The medical world is asking questions regarding head injuries (concussions) and long-term neurological disease. What relationship could concussions have with Chronic Traumatic Encephalopathy or Alzheimer's. The death of several well known athletes, in particular Junior Seau, the Hall of Fame football player, brought to light the issue and focused more attention on it. As did the need to gather data on head impacts, and other data to see if data and technology can help provide meaningful information to the medical community, and hopefully understand better the causes and become pro-active in limiting injuries and neurological diseases related to the injuries.

KEYWORDS

i523, hib315, Big Data, Concussions, Traumatic Brain Injury, Data Types, TBI, Data Modeling

1 INTRODUCTION

"I'm back you're home the days you really miss me. I guess you did by the look in your eyes. Now lay back and relax let your body put away the distance then you and me can rock a bye." So says Anita Ward in her classic 1979 disco anthem - Ring My Bell. Back in the day getting your "bell rung" meant something quite different and much more serious.

The US National Library of Medicine defines concussion as a minor traumatic brain injury (TBI) that may occur when the head hits an object or when a moving object strikes the head. Typically with a TBI, there is leakage from a blood vessel in the brain as a result of the trauma within the brain. This leakage can accumulate in the skull and the resulting pressure can lead to brain damage and even death.

"In 2015, approximately 2 million individuals suffered with TBIs in the USA alone, and the number worldwide was approximately 60 million. The medical, economical and social expenses directly related to TBI are approximately 96 billion dollars annually in the USA alone. Injuries that include TBI cause the deaths of approximately 150 people per day in the USA resulting in approximately 50,000 deaths per year."[5] Even more concerning, "nearly one-third of athletes have sustained a concussion that went undiagnosed and risked further brain injury, according to the Clinical Journal of Sport Medicine" [1]

With the increase in technology, i.e., smart helmets, helmet inserts, and data gathering; can Big Data play a role in keeping our athletes healthier? The National Football League thinks so, as they have hired Biokinetics Inc. to provide data from the testing of 17 helmets used in the 2015 football season. "The National Football League has funneled millions of dollars into big data and biosensors to help understand player injuries better."[2]

2 THE TECHNOLOGIES

The science behind TBIs as well as the technologies available to athletes has grown significantly. For example, helmet manufactures have various types of "smart helmets". University of Wisconsin students are developing a football helmet with "brain wave probes and a device that measures acceleration forces to detect concussions on the field and directly communicate the information to medical staff". [4] Riddell, a leading US football helmet manufacturer has the InSite system that gathers data via sensors and communicates wireless alerts and information to medical staff and coaches real-time. Numerous manufacturers have some version or variation of the *smart helmet*. Most however measure the impacts or force of the impacts and gather that information. Over the last several years the sensors have become more accurate in measuring force detail, like location and g-force (compared to the force of gravity).

Unequal Technologies has designed a helmet liner that fits over the existing helmet. One of the most interesting approaches to helmet development is that by Colorado Springs engineer Troy Fodemski. Mr. Fodemski's version of the *smart helmet* imagined a "helmet use of sensors to measure a hit, compare it to a set of criteria, and deploy up to 75 airbags inside the helmet that would precisely cushion the area of impact, thereby stopping the brain from forward movement. This helmet is the first to use airbag technology."[6]

3 THE DATA

With all this data from the helmets now the data scientist is ready to go, right? If only it were that simple. Data gathered from the helmet is actually just one type of data but there are many more. It's important to note that the study of TBIs is an ongoing effort with ongoing discoveries. The effort is to understand TBIs, understand the causes, symptoms and try to limit or eliminate them. Therefore, understanding historical research, similarities to animals and the corresponding historical research, biomarkers, neurological changes, imaging, sensors and other medical related research are all data types that play a role.

This is a good time to mention the three V's of big data: Volume, Variety and Velocity. With all of these data types, along with the amount of data needed to support high definition imaging, the recording of daily activity with the use of the helmet, historical data and the like, the volume can be very large. The data variety comes in the form of structured and unstructured data, anything from historical medical research documents to measured sensor numbers, technical chemical measurements and images. Data velocity is diverse with some data arriving daily or hourly (*real time*), while access to medical research journal is based on accessibility and download speed.

4 HISTORICAL MEDICAL RESEARCH

There is extensive historical medical research into TBIs, much of which is tied to neurological measurements with the patient - *after the fact*. It's also from this research that we are beginning to learn the long-term affects of TBIs on the health of the athlete. Additionally, some related research has been done on animals, but making a correlation between the animal and a human can be problematic. For the data scientist, most of the research is high quality but the conclusions can be challenging to manage. Comparing them to other data sources and making them relatable across data types to the data scientist is the problem.

5 BIOMARKERS

To the data scientist, this data may need to be broken down into more granular detail like, neurological measurements (behavioral, neurological function like memory or mood, and items that are more objective to measure) or psychological measurements. However, we know "Tau is a protein that forms in the brain when someone experiences a concussion, or any form of brain damage.".[1] The measurement of cerebrospinal fluid and blood can be used to measure changes in proteins as well as other biomarkers. There is extensive research on biomarkers for TBIs, which is good news for the data scientist, something quantitative that can be used. However, most of these biomarkers are based on severe TBIs. "In mild TBI/concussion where imaging is negative, there is a substantial need for blood - or CSF-based biomarkers. Also, even though current blood-based biomarkets can indicate the extent of damage, they do not provide information about the pathological changes of the secondary injury process, and thus they cannot identify therapeutic targets or help with evidence-based therapy."[5]

6 IMAGING

The good news for the data scientist is that this is enriched data that has the promise of a level of consolidation and is quantifiable, that is, the data can be managed. There are imaging tools available like Brain-Map and others that can help with the mining of data; however, challenges abound for the data scientist due to data size, lack of agreed upon and consistent methods, and standards. A standard MRI is about 5-6 MB's while, "just the acquired neuroimaging data alone are an average of 20 GB per published study." [5] The expectation is that the file size for imaging will double every two years. And much like biomarkers, imaging provides more definitive data for the serious TBIs versus the more mild episodes. "The main challenge is standardization or how to take into account differences between various laboratories using different acquisition rates, resolutions, scanning parameters, among others."[5]

7 SENSORS

TBIs use to be seen based on those that had a clear neurological impact. On the football field we would here "his bell was rung", based on the fact the athlete would briefly lose consciousness or memory. However, current medical thought is that it may not always be the big hit, the *bell ring* that causes the long term damage but milder TBIs that happen more frequently and over a longer period of time.

The key then is to try to "measure" the TBIs. "Because TBI is caused by physical forces that can be measured and quantified, an important step toward using BD (Big Data) approaches and developing a TBI dosimetry is to understand the correlation between the physical forces and the biological response."[5] "Thus, approaches to track and gauge the cumulative effects of repeated mild TBI are at the forefront of investigation. Understanding the relationships among frequency, location, force and thresholds for concussion with those of acute and long-term changes in physiology, cognition, vision, balance and presence of blood markets is hindered by accuracy of recording impacts sustained."[5]

Sensors can vary in function but three agreed upon variables that need to be captured are the number and location of the head impacts along with the measurement of g-force. The number, location and force of the hit is critical to bridging the gap between the physical force (the hit) and the biological response.

8 WHAT'S A DATA SCIENTIST TO DO

The data has to be collected at the most granular level. At this point there will need to be some extensive modeling done to start to build relationships between the different types of data. This includes the critical step of building a correlation between the data types. For example, given physical force of A and a brain image within category B, we see the biological response of C. From here we can build additional models and change existing ones as the data builds, and we increase our ability to quantify variables, making some unstructured data structured and moving from big data to smart data. "There is a degree of agility and flexibility in the sort of modeling required for Smart Data that vastly exceeds that of non-Semantic data. In the case of the latter, Data Modelers have to determine in advance of the model's creation each and every question that the model will answer, and how it is going to relate to specific facets of known data types. Such a process is not only arduous and type consuming, but makes it difficult to add new sources or to change the requirements for a model."[3]

9 CHALLENGE AND CONCLUSION

Ironically, when a boxer is knocked out, the ring-side bell is actually rung. Any athlete, be it a football player, hockey player or boxer needs big data. Not only can big data help with gleaning meaningful information from a plethora of related sources, it has to! With the technology to gather and manage data in data lakes, and with intelligence garnered from detailed data collection, and modeling, it's big data that can provide the wisdom needed by the medical community. It's big data that can ring the bell and help stop the progression of these diseases and keep our athletes healthier.

ACKNOWLEDGMENTS

Many thanks to Professor Gregor von Laszewski, the Teaching Assistants and Indiana University. I also want to thank Katie, my understanding wife. Lastly, for my employer AT&T for a commitment to education and giving me 26 years of experience, challenge and opportunity.

REFERENCES

[1] Admin. 2016. Med Students Build Smart Helmet to Detect Concussions. Online. (Aug. 2016). http://eptechview.ttuhsc.edu/ttuhsc-el-paso/med-students-build-smart-helmet-to-detect-concussions-2/ Texas Tech University Health Sciences Center El Paso (administrative post).

[2] Jordan Davidson. 2015. NFL taps big data to study concussions, but major game changes far off. Online. (Aug. 2015). http://www.zdnet.com/article/nfl-big-data-concussions-innovation-results-a-way-away/ Published for Between the Lines.

[3] Jelani Harper. 2015. The Evolution of Big Data to Smart Data. Online. (May 2015). http://www.dataversity.net/the-evolution-of-big-data-to-smart-data/ via Dataversity website.

[4] Karen Herzog. 2014. UW student use 'smart' technology in football helmets to detect injuries. *Journal Sentinel (Milwaukee - Wisconsin)* (Dec. 2014). http://archive.jsonline.com/news/education/uw-students-use-smart-technology-in-football-helmets-to-detect-injuries-b99406965z1-286178071.html

[5] Denes V Agoston & Dianne Langford. 2017. Big Data in traumatic brain injury; promise and challenges. (May 2017). https://www.futuremedicine.com/doi/10.2217/cnc-2016-0013

[6] Natasha Townsend. 2013. Engineering and Concussion Reduction Therapy (CRT). online. (Feb. 2013). http://www.designworldonline.com/engineering-and-concussion-reduction-therapy-crt/ via Design World website.

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 artifical 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 achievment 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 conncected 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 perceptable 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, unecessary expenses and unecessary 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 appropirate 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 proccessing 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 andXueChen 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 Gargeya, 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

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 [13]. 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 [8]. 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.

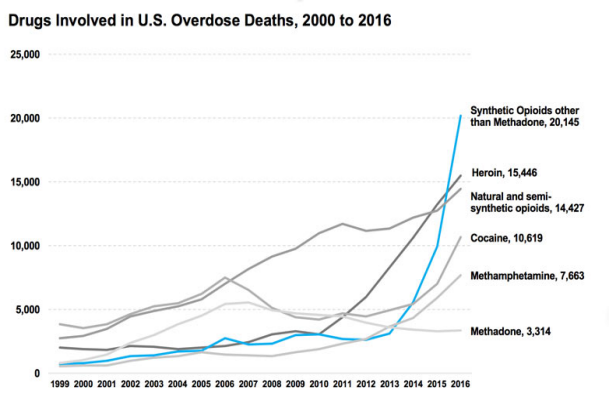


Figure 1: Drugs Involved in U.S. Overdose Deaths from 2000 to 2016, National Institute on Drug Addiction (NIDS) [14]

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 personal 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

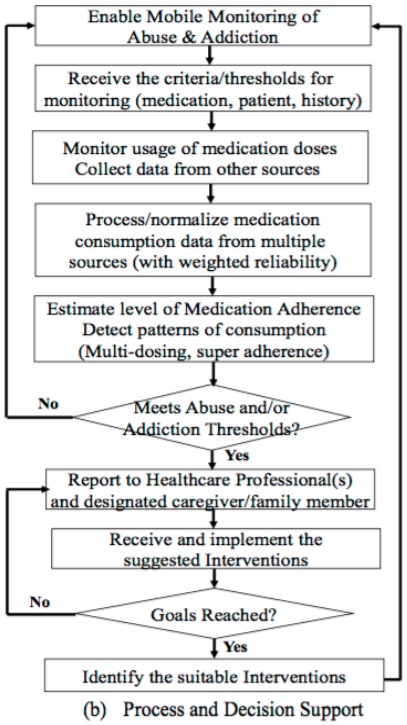
Technology-based interventions have been used for drug addiction assessment, treatment, prevention, and recovery [11]. In terms of assessment, data about substance use can be obtained from mobile cell phone reporting outside of treatment settings. Web-based approaches to treatment have been implemented online to improve behavioral and psychosocial functioning for addicted individuals in recovery [12]. 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 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-occurence, 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 [10]. 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 [9]. 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 [10].

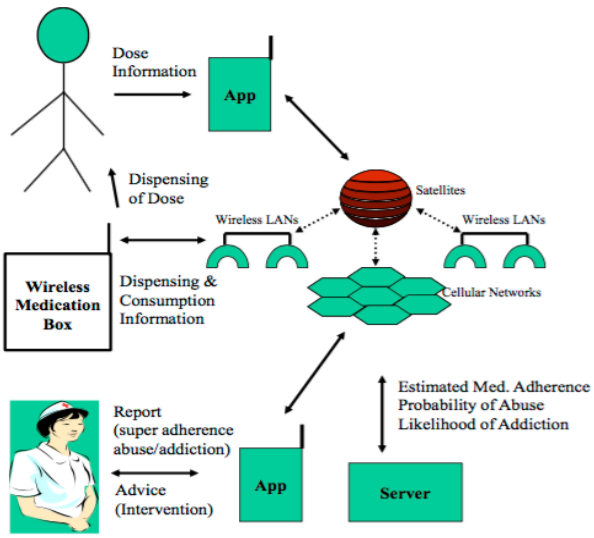
1.3 Medication Adherence and Abuse Monitoring System

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



(b) Process and Decision Support

Figure 2: Process and Decision Support for Abuse Monitoring System [19]



(a) Architecture of the Abuse Monitoring System

Figure 3: Architecture for Abuse Monitoring System [19]

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.

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.

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 [7]. 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.

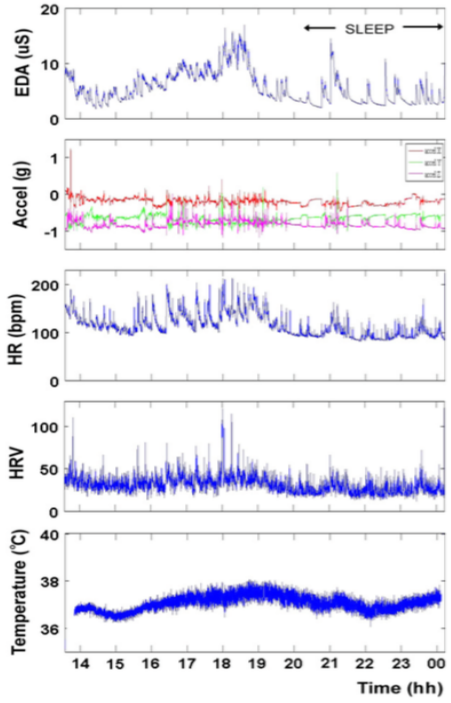


Figure 4: Sample Ambulatory Data from Rhesus Macaque Recorded on Wearable Sensor for 11+ hours Inside Large Primate Cage Facility [7]

1.5.1 LoRa Backscatter: Enabling Ubiquitous Connectivity. Emerging technologies, such as longe 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.

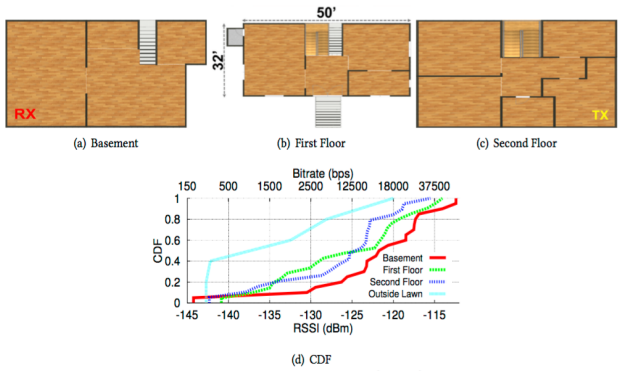


Figure 5: Home Deployment of LoRa backscatter packets across 4,800 sq. ft. House Spread Across Three Floors [17]

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

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 kpbs	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 kpbs	yes	yes	yes	yes

Figure 6: LoRa Backscatter Epidermal Patch [17]

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 (ECK), 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.

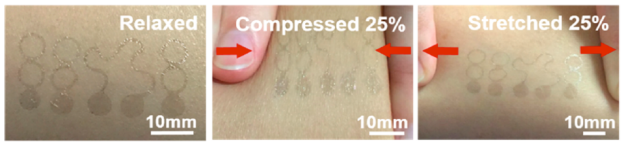


Figure 7: Graphene Electronic Tatoo Biosensor [1]

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] Edward W. Boyer, David Smelson, Richard Fletcher, Douglas Ziedonis, and Rosalind W. Picard. 2010. Wireless Technologies, Ubiquitous Computing and Mobile Health: Application to Drug Abuse Treatment and Compliance with HIV Therapies. *Journal of Medical Toxicology* 6, 2 (July 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), 881fi?!8. https://doi.org/10.1176/appi.ajp.2008.07111835

[7] Richard Ribn 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, 4046–4049. https://doi.org/10.1109/EMBC.2012.6346855

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

[9] 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

[10] 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.* 13, 5 (2011), 390–397. https://doi.org/10.1007/s11920-011-0218-y

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

[12] 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

[13] David Metcalf, Sharlin T. J. Milliard, Melinda Gomez, and Michael Schwartz. 2016. Wearables and the Internet of Things for Health. *IEEE Pulse* 7, 5 (Oct. 2016), 35–39. https://doi.org/10.1109/MPUL.2016.2592260

[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* 18, 2 (June 2016), 213–221. 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). arXiv:1705.05953 http://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/NEJMp1402780 arXiv:http://dx.doi.org/10.1056/NEJMp1402780 PMID: 24758595.

How the Datafication of Activity is Improving Human Health

Ross Wood
HID345
rmw@indiana.edu

ABSTRACT

As the world becomes more technologically advanced, more and more work is being done on computers digitally versus in the real world physically. This shift in how work is done is creating a situation where huge swaths of people are sitting down for longer periods of time than is medically recommended for healthy living. For many, this has created the need for more exercise and athleticism than what one typically gets in a normal day. This occupational atrophy is one of the driving forces in the development of technology that monitors and generates data of the user's day-to-day physical activity and vital signs. All this new data being created is helping make improvements in the lives of those with sedentary jobs, while also revealing new techniques and applications for already existing training or exercise programs. The data is being generated at an increasing rate from the growing number of users who track their activity with the use of sports trackers, smart clothing or athletic wear, and mobile phones. The applications of this data are not limited to improving the health of office workers and other employees whose jobs or lives require low physical activity. The datafication of physical fitness and activity is also changing the way professional athletes and coaches approach their training regiments, as well as having military applications in regards to training personnel.

KEYWORDS

i523, HID345, Fitness Tracker, Data Science Exercise, Big Data Exercise, Smart Clothing Data, Activity Data, Smart Clothing

1 INTRODUCTION

The big changes taking place as work and labor evolve with technology in the modern world are part of the reason why sports trackers, smart clothes, and mobile health apps are gaining in popularity. For someone who has to work at a computer for six to eight hours a day, a digital reminder to be active, the ability to monitor heart rate, caloric intake/output, or even just steps, can create a huge difference for the individual's health. For people who find themselves in jobs that are not physically demanding and who want to build their athleticism and improve their health, sports tracking and the data they generate can be a good way to stay on track. Through the analysis of personal data generated through activity, users can track their progress, discover and improve areas of difficulty, and reach their exercise goals more easily and efficiently. The technology being created to achieve this, and subsequent data that is generated from it's use, is good for more than just helping people who want to get into shape achieve there goals. All this new technology and the benefits that come with analyzing the data it produces is beginning to be seen in every faucet of society. It is affecting everything from the world of professional sports, to nursing homes, to having some military applications. A new age of personal health is being ushered into reality thanks largely to the ability to create, track, monitor, and analyze all this new activity data.

2 USEFULNESS OF SPORTS TRACKERS

Sports trackers are tools that users wear to monitor their activities. Colloquially referred to as a Fitbit, these devices are anything that is worn externally that creates data from tracking activity. More and more mobile phones are now coming with standard hardware that enables basic activity tracking and users are taking advantage of these updates by downloading and using apps to track their activity as they progress towards their exercise goal. Smart clothing is a growing field that is presently in its infancy. Smart clothes are any clothes that have sensors built into them that monitor and creates data on human activity and can monitor vital signs. These blooming technologies are paving the way towards a new understanding of personal health.

Studies have shown that activity tracking technology has proven to increase activity in adults who are overweight or would be classified as obese [3]. By reinforcing healthy habits and activities, the sports trackers, apps, or clothes help to encourage a routine of healthy living. Other features, like social media networks associated with tracking apps, have also proven to have a positive influence on user activity [7]. With easy access to social sports knowledge and these activity tracking tools, humans are entering into a personal health era the likes of which have never been seen before.

3 HOME CONSUMER HEALTH BENEFITS

The average consumer working a sedentary job is one of the groups benefiting most from the adoption of technologies that track and record all the activity data one creates throughout the day. Be it a wrist sport tracker, mobile phone, or a new article of smart clothing, an individual having access to the activity data they generate is important because they can use it to monitor their health, improve their lives, and achieve their fitness goals. With access to their own data, the home consumer gets to become a junior data scientists as they analyze their activity data to improve their performance. The constant ability to assess and monitor their progress and performance is allowing users to shorten plateau time and reach their goals even faster.

The amount of data being generated by active users is growing at an exponential rate as more consumers begin adopting these activity tracking devices. In December of 2015, it was estimated that the amount of new data being generated every day from athletic wear, phones, or trackers "can easily reach billions of tuples per day" [2]. This number is going to continue to rise as the technology improves. The idea of athletic wear via smart clothing is still more or less in in its infancy, but once it starts to become more popular

and cost efficient, the amount of data generated is going to start increasing at an even faster rate.

Indeed, the amount of new data being created every day is showing no signs of slowing down. The smart clothing market, which has been steadily growing since 2015, is expected to overtake mobile phones and sports trackers in generating activity data [1]. Since "clothes outsell phones"[1], this data explosion and all the benefits that come with it are only going to continue to boom in the coming decades.

3.1 Senior Benefits

These technologies are doing more than just benefiting those who want to achieve greater athleticism or get into shape. They are also helping seniors approach individual personal health in their twilight years in ways no generation has been able to in the past. With access to trackers that monitor their fitness levels and up to the minute details like heart rate, steps, or calories burned, seniors are better able to monitor their own health. This ability to monitor their health has a cascading effect in that better health and fitness has other benefits like stronger bones and better coordination, which can prevent falls and other hazards [9]. As medicine and advances in health care continue to improve and the human lifespan gets longer and longer, these technologies are proving to be a major benefit to groups whose health is at the greatest risk.

Right now, a reticence to adopt new technology is one of the biggest setbacks preventing seniors from taking advantage of activity tracking and the benefits that come with it. This is primarily caused by how quickly technology changes and difficulties learning a new technology. As time goes by and younger generations who have grown up with hand held devices and sports trackers become older, this setback will become less of a problem [9]. As individuals exercise greater control over their own health with help from these tools, humans are going to continue to prolong their lifespans, further shortening the divide between number of years and number of healthy, active years.

3.2 Social Media User Benefits

Another group that benefits from the use of this technology is social media users. Online activity and athletic social media communities are one of the fastest growing online social media communities [7]. Research has found that creating a social network around sports tracking is a novel approach to motivating users to engage in activity more, worry more about their diets and physical condition, while also producing, accessing, and learning important sports knowledge from the online sports community [7]. Since the primary way to interact with online friends on a social network built around tracking your activity is to be active, this again contributes to the cascading effect in that people will want to exercise more to interact with their online friends more. This is motivating people to push themselves harder and achieve greater levels of physical health.

3.3 Athletics Organization Benefits

Health and fitness are not the only areas where this newly created activity data is helpful. Sports organizations from the high school level to professionals are improving their teams' performances from analysis of this data. The technology to track an individual athlete's performance is getting more advanced. "Sensor technology in sports equipment such as basketballs or golf clubs also allows us to get feedback on our game . . . using smart technology to track nutrition and sleep, as well as social media conversation to monitor emotional well being" [6]. This new approach to monitoring athletes and their activity is more complex, drills deeper, and is proving to be vastly superior to previous methods of statistical analysis used to improve performance [6]. Starting in the early 2000s, teams that have begun taking advantage of these techniques have all improved their performances.

3.4 Military Benefits

Using many of the same techniques that athletic organizations are using, militaries around the world are beginning to adopt a lot of the same approaches when training their soldiers. As technology that works in sync with human activity becomes more advanced, so too does the modern soldier's reliance upon it in order to maintain a combative edge. Smart clothing that can monitor for potentially hazardous environmental conditions, as well as physical and mental health and safety are among some of the technologies being developed that assist the user and create data from monitoring human activity [8]. The United States military is dedicated to the development of these smart clothing technologies that will benefit their soldiers [4]. An example would be an article of clothing that can detect an injury and automatically call for help, or even administer some rudimentary medical treatment to help save the solder's life. As is often the case, the development of these technologies frequently find their way to the private sector and civilian use. Thus, even more data and technology to monitor human activity is on the horizon.

4 PRIVACY CONCERNS

As is usually the case with all the data creation in the 21st century, it has the potential to be a double edged sword. While undoubtedly useful to the individual user and society as a whole, the data created from sports trackers, smart clothes, and mobile phones are also readily available to the technology and software creators [5]. An example of how this data could be used against the user is through sharing the user data with other organizations. One example of how this data could hurt the user would be if an individual's personal heath or activity information were somehow made available to another companies who could potentially gouge their customers based on unhealthy activity or any other number of variables. Even though the data is private, it is possible to use machine learning techniques to identify users around the world. There are few modern laws that govern what companies and organizations are allowed to do with their user generated data. This type of privacy concern is just now beginning to be addressed in courts around the world, but for now, it remains a valid point of concern for the ever growing number of users.

5 CONCLUSION

Fitness tracking technology is making it easier for humans to achieve their fitness goals in ways they have never been able to in the past. With new research and development being done everyday to improve the already existing tools and techniques, the

technologies are only going to continue to get more precise and efficient in regards to helping humans monitor their activity and quantify their vital signs. The myriad of organizations discovering the usefulness of this approach to activity data analysis is pushing this field to greater heights every year. Monitoring and studying our individual activity data is shaping up to be the key to having a long and healthy life.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Miao Jiang, and Juliette Zerick for assistance with this assignment and using github.

REFERENCES

[1] Laura Bellamy Pauline Burke Rajiv Ramanathan Vijay Balakrishnan Alex Hanuska, Bharath Chandramohan. 2015. *Smart Clothing Market Analysis.* Technical Report. Sutardja Center for Entrepreneurship and Technology,.

[2] Rudyar Cortés, Xavier Bonnaire, Olivier Marin, and Pierre Sens. 2014. *Sport Trackers and Big Data: Studying user traces to identify opportunities and challenges.* Research Report RR-8636. INRIA Paris. https://hal.inria.fr/hal-01092242

[3] Herman de Vries, Thea Kooiman, Miriam Van Ittersum, Marco van brussel, and Martijn Groot. 2016. Do Activity Monitors Increase Physical Activity in Adults with Overweight or Obesity? A Systematic Review and Meta-Analysis. *Obesity, A Research Journal* 24 (09 2016), 2078–2091.

[4] Paul B. Lester, Sharon McBride, Paul D. Bliese, and Amy B. Adler. 2011. Bringing science to bear: An empirical assessment of the Comprehensive Soldier Fitness program. *American Psychologist* 66, 1 (2011), 77–81. https://doi.org/10.1037/a0022083

[5] Chantal Lidynia, Philipp Brauner, and Martina Ziefle. 2017. *A Step in the Right Direction – Understanding Privacy Concerns and Perceived Sensitivity of Fitness Trackers.* Springer International Publishing, Cham, 42–53. https://doi.org/10.1007/978-3-319-60639-2_5

[6] Bernard Marr. 2015. *Big Data - Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.* John Wiley & Sons Inc, Hoboken, New Jersey. http://www.ebook.de/de/product/23189364/bernard_marr_big_data_using_smart_big_data_analytics_and_metrics_to_make_better_decisions_and_improve_performance.html

[7] Jarno Ojala and Johan Saarela. 2010. Understanding social needs and motivations to share data in online sports communities. In *Proceedings of the 14th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '10.* ACM Press, Tampere, Finland, 95–102. https://doi.org/10.1145/1930488.1930508

[8] Sofia Scataglini, Giuseppe Andreoni, and Johan Gallant. 2015. A Review of Smart Clothing in Military. (05 2015).

[9] Trinidad Valenzuela, Yoshiro Okubo, Ashley Woodbury, Stephen R. Lord, and Kim Delbaere. 2016. Adherence to Technology-Based Exercise Programs in Older Adults. *Journal of Geriatric Physical Therapy* 60 (jun 2016), 151–156. https://doi.org/10.1519/jpt.0000000000000095

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 learning some advanced information of this game by researching win rate and something else. These could help players getting a better performance in the game. Further more, 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 succeed. 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. Further more 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 match 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]

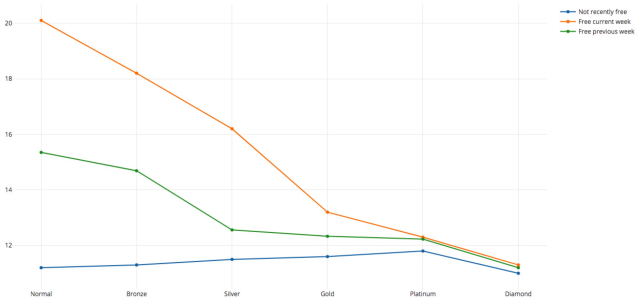


Figure 1: Usage of free champion

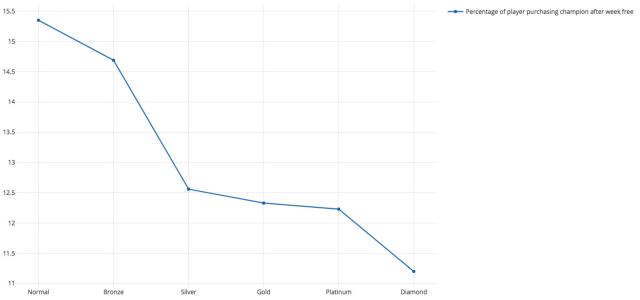


Figure 2: Probability of purchasing champion

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.

As the data from Figure 2, we can get the probability of players who bought champions after playing with free champion rotation.

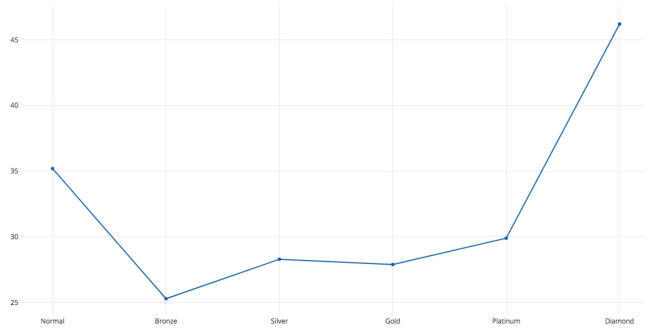


Figure 3: Percentage of players willing to buy skins

And we can get:

$$I = \sum_{c=1}^{10} \sum_{t=1}^{6} a * U * p_c \tag{2}$$

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.

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.

$$I_s = \sum_{c=1}^{10} \sum_{t=1}^{6} b * U * p_s \tag{3}$$

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 means 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 got 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.

$$KDA = \frac{Kill + Assist}{Death} \tag{4}$$

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.

$$T = \frac{KDA * cs}{R} \tag{5}$$

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).

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 study provided a short overview of disability informatics, providing examples of using health informatics for people with disabilities.This study also explored the potentials of using big data sources to better understand people with disabilities and uncovered the potential of big data and cloud computing in developing assistive technology and information technology for people with disabilities.

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 UTHealth, 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 [6]. Notably, the number of people with disabilities is expected to increase as a result of extension of human life-span, decreases in communicable diseases, the improvement of medical technology, and decrease of child mortality [13]. According to United Nation, people with disabilities are the largest minority group in the world [1, 4, 6]. 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 [6]. Population aging trend also signifies that disability will be a more common and urgent issue in the future [7]. 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, gender and social and cultural factors) for that individual". While people with disabilities are those people who have limitations in their actions or activities resulting from physical or mental impairments, however, there are many types and levels of disabilities and their actions and activities are affected differently by their disabilities [1]. The complexity of disabilities presents difficulties and challenges to accommodate the different needs of people with disabilities and improve their qualities of life [11].

The development of digital technology has changed many people's lives, the life of people with disabilities has also been improved by technology [11]. People with poor visions can using cell phones to contact others, access information online with screen readers. People with hearing problems can text other people with their cell phone. This study is trying to help people with disabilities from a health informatics prospective, specifically the disability informatics, and looking into how big data and cloud 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 [13]. 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) [12]. Braille has been traditionally used by people with visual impairment, but now was replaced by technology such as voice synthesis and screen readers [12].

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 of health informatics that is defined as "any application that collects, manages, and distributes information that are related to people with disabilities, as well as to care givers (including familiar members and health care providers) and rehabilitation professionals" [1]. Disability informatics is closely related to other health informatics areas such as medical informatics, public health informatics and consumer health informatics, because people with disabilities usually have some secondary medical condition such as poor health status and increased personal health care needs. Gather medical and health information

can help to better understand and accommodate people with disabilities [12]. A study from the early 2000 has identified the potential of public health informatics for prevention at all vulnerable points in the causal chains leading to disability and proposed that applications should not be restricted to particular social, behavioral, or environmental contexts, but in a more global context [15]. Another previous research has designed and deployed an extended version of Artemis system (a cloud system designed to acquire data and store physiological data of clinical information for real-time analytics) in a hospital. They have identified that high speed physiological data produced at intensive care units as big data, and the proper use of such data can promote health, reduce mortality and disability rates of critical condition patients and create new cloud-based health analytics [9]. Research also has shown that many disabilities are genetic, therefore, bioinformatics has implications in the education of genetic screening and gen therapy treatments in the future [1]. People with disabilities usually need some assistive technology in their daily life. These technology that assist them to perform basic physical and social functions. The use information technology and assistive applications in disability informatics are categorized into three areas: virtual, personal, physical.

3.1 Virtual Environment

The digital revolution had and will continue to have a profound positive impact on the life of people with disability by empowering them with the help of digital technologies [1]. However, there are still access issues in the digital world. One of the barrier is the use of the World Wide Web (WWW or Web). The WWW has always had a strong awareness and been advocacy for accessibility since early on in its evolution. The World Wide Web Consortium (W3C) had passed the Web Accessibility Initiative (WAI) and Web Content Accessibility Guidelines in the late 1990s [1]. A number of assistive technologies were designed to help people with disabilities to use the Web. For example BBC Education Text to Speech Internet Enhancer (BESTIE) is a CGI Perl script that can help people with disabilities who are using text-to-speech systems for Web browsing to modified the web page removing images, Java and Javascript code that may cause difficulties to understand the BBC web page content [5]. However, the limitation of BESTIE is that it is only compatible with BBC website. Other researchers also came up with Personalizable Accessible Navigation (PAN), which is a set of edge services designed to improve Web pages accessibility which allow personalization and the opportunities to select multiple profiles, making it compatible for web as well as mobile devices [10].

3.2 Personal Environment

Disability informatics also emphasis on providing safe personal environment for people with disabilities, Health monitoring is a very important area. Technologies like small tracking device can monitor heart rate, blood pressure, also allow people to call for help easily and smart clothing and even smart furniture have been developed to monitor people's health status and can help provide people with disabilities a safer personal environment and also provide health information for their medical care providers [1]. However the ethic of such health monitor devices are always in debate, some believe it can be an invasion of privacy and a restriction of personal freedom, others hold the ground that its main purpose is to help people with disease or disabilities, since it can alert their caregiver if the individual are exposed to harm (such as a person with mental disability and has a history of self-harming, these device can prevent unwanted behavior) [3].

3.3 Physical Environment

Since the American Disabilities Act passed in the 1990s, the accessibility of physical environment has been improved in a great degree. However, people with disabilities still would meet some barrier and problem, one of them is the lack of curb cuts. Assistive information technologies has been developed in an effort to solve this problem. One of them is MAGUS, which is a project using geographical information system to inform users about wheelchair accessibility in urban areas [1]

4 BIG DATA AND CLOUD COMPUTING FOR DISABILITY INFORMATICS

The contribution of Big data and cloud computing have been recognized and accepted by researchers in health informatics [14]. The potential of big data and cloud computing for disability informatics and for people with disabilities has been explored by a few researchers and organizations. Data-Pop Alliance is one of the organization has recognized the big data and potential for study and help people with disabilities for disability informatics and people with disabilities [11]. Their research has categorized three type of big data source used across disability research: exhaust data (mobile-based data, financial transaction, transportation and online trace), digital content (social media and crowed-sourced/online content), and sensing data (physical and remote) [11]. They also provided the potential for some of these data sources, for example, researchers can use transaction data to compare cost, availability, and use of services that offer accessible options (such as accessible hotel listings) [11]. They also suggested that researcher can use social media data to represent people with disabilities as a network of interaction and using crow-sourcing to map the locations of accessible businesses and public places [11]. The organization has also identify four functions of big data on disability:descriptive, predictive, diagnostic and engagement. Descriptive function of big data is to describing and presenting the collected information such as using location data to map workplaces that are accessible to people with disabilities [11]. Predictive function is making inferences based on collected information such as discovering trends in the growth of number of accessible businesses in a certain urban area, while the diagnostic function means establishing and making recommendations on the basis of causal relations such as showing what can help increasing accessible business in a certain area [11]. Finally, the engagement function refer to shaping dialogue within and between communities and with key stakeholders through communication of data [11].

Cloud computing in combined with big data can also provide great opportunities for research and improvement of quality of life for people with disabilities [2]. The term cloud "refers to everything a user may reach via the Internet, including services, storage, applications, and people" [8]. Depending on the type of using, the "cloud" can be use for different purpose, such as for companies, the cloud could be used for hosting services so as to avoid the costs and

difficulties associated with hosting onefis own servers and software and for individuals, the could is often used as information storage [9]. Regardless of the types of usages for cloud, the end using must still access the information and services residing in the cloud through device like a smart phone or computer [8]. Cloud computing has been used to provide more accessible virtual environment, especially Web access through project like WebAnywhere, which is a cloud based tool for blind using to access Internet [8].

Cloud computing and big data analytics can also be helpful in health monitoring. The Artemis project mention earlier provide a example of big data analytics and cloud computing usage in health monitoring, by creating new cloud-base health analytics solutions [9]. Previous researchers have developed a mobile app to collect motion data of Parkinson's disease (PD) which is a disease resulting in mobility disorder using the smart phone 3D accelerometer and to send the data to a cloud service for storage, data processing, and PD symptoms severity estimation, which provide an user-friendly and economically affordable system to monitor and assess the condition of PD [10]. Although this system is not for people with disabilities, but it provided potentials for similar systems to be developed for different kind of disabilities.

Another application of cloud computing and big data in assistive technology is the CloudCast platform, which is a cloud-based speech recognition services that can be used for many assistive technology application for people with speech difficulties and hearing impairment, it also facilitate the collection of speech data required for the machine learning techniques [3]. Similar to Alexa Voice Service, it provide reliable speech recognition which can be used with assistive devices for people with hearing impairments, but CloudCast platform also provide customization for assistive technology applications benefiting users with speech impairment [3]. This research provided a great example of using big data and cloud computing in combine to solve a certain problem for people with disabilities (in this case it is barriers for speech impairment).

5 CONCLUSION

People with disabilities has long been considered as underprivileged groups. Although the development of information technology and assistive technology has improved the life of people with disabilities, there is still much left to done. This study provided a short overview of disability informatics, providing examples of using health informatics for people with disabilities. This study has also find out there are great potentials to use big data source to better represent people with disability and identity and study issues and propose actions and solution to the challenges faced by people with disabilities. Cloud computing and big data can also help improving assistive and information technology that are now used to help people with disabilities. However, there are still a lot challenge faced by researchers and organizations interested in improving the quality of life for people with disabilities. The most dominated challenge is the different needs for people with different disabilities types and function levels.

The limitation of this study is that there are limited number of studies on disability informatics or big data. This study draws some of the examples from health informatics which their study focuses were other disease. However, these examples do show opportunities to developing similar systems for special needs of people with disabilities. Another limitation is that also there are organizations that have illustrated the potentials to use big data in analyzing and understanding people with disabilities, there are not yet studies has been done to prove these potentials. However, this leaves opportunities for future researchers to use big data to better understand the needs and behaviors of people with disabilities.

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

[2] Ann Cameron Caldwell. 2011. *Untapped Markets in Cloud Computing: Perspectives and Profiles of Individuals with Intellectual and Developmental Disabilities and Their Families.* Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Chapter 30, 281–290. https://doi.org/10.1007/978-3-642-21663-3-30

[3] Stuart Cunningham, Phil Green, Heidi Christensen, JJ Atria, A Coy, M Malavasi, L Desideri, and F Rudzicz. 2017. Cloud-Based Speech Technology for Assistive Technology Applications (CloudCAST). *Harnessing the Power of Technology to Improve Lives* 242 (2017), 322.

[4] 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

[5] Ugo Erra, Gennaro Iaccarino, Delfina Malandrino, and Vittorio Scarano. 2007. Personalizable edge services for Web accessibility. In *Universal Access in the Information Society (W4A)*, Vol. 6. WWW2006, ACM, Edinburgh, UK, 285–306.

[6] Lex Frieden. 2015. Why Disability Informatics? (02 2015). https://sbmi.uth.edu/blog/feb-15/021115.htm

[7] Jan Grue. 2016. The social meaning of disability: a reflection on categorisation, stigma and identity. *Sociology of Health and Illness* 38, 6 (2016), 957–964. https://doi.org/10.1111/1467-9566.12417

[8] Jeffery Hoehl and Kaleb August Sieh. 2010. *Cloud Computing and Disability Communities: How Can Cloud Computing Support a More Accessible Information Age and Society?* Technical Report. Silicon Flatirons Center, Colorado, US. https://doi.org/10.2139/ssrn.2285526

[9] 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

[10] Di Pan, Rohit Dhall, Abraham Lieberman, and B. Diana Petitti. 2015. A Mobile Cloud-Based Parkinson's Disease Assessment System for Home-Based Monitoring. *JMIR mHealth uHealth* 3, 1 (26 Mar 2015), e29. https://doi.org/10.2196/mhealth.3956

[11] 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.

[12] Paraskevi 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

[13] 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

[14] M. Viceconti, P. Hunter, and R. Hose. 2015. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics* 19, 4 (July 2015), 1209–1215. https://doi.org/10.1109/JBHI.2015.2406883

[15] WA Yasnoff. 2000. Public health informatics: improving and transforming public health in the information age. *Journal of public health management and practice* 6, 6 (11 2000), 67–75.

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 dataset 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 dataset.

ID3 algorithm is well known as Iterative Dichotomiser 3 algorithm invented by Ross Quinlan. It is used to generate a decision tree recursively from a dataset. Then, the output decision tree could be used as a model to predict (classify) any input instance as which class or group it belongs to. The ID3 algorithm starts at the original dataset as the root node. Then in the iteration (recursion), the algorithm calculates the entropy of the label and the information gain of each unused attribute. The algorithm selects the attribute which has the greatest information gain and separates the dataset into multiple subset according to possible value of this chosen attribute. Each subset of the original dataset is an inner node in the final decision tree. ID3 algorithm continues to call itself (recursion) until two conditions are satisfied. First condition is that every element in the subset belongs to the same class, the subset denoted as the leaf of the decision tree is marked as the name of that class. Second condition is that there are no more attributes to be selected, but the examples still do not belong to the same class, then the node is turned into a leaf and labelled with the most common class of the examples in the subset[4].

The concept entropy this paper discussed above is derived from Physics, which is a measure of how much chaos or uncertainty of the system. In big data, it refers to the amount of uncertainty of the given dataset. Entropy is usually represented by H(S). It could be calculated from the probability of each element in the label of a dataset.

$$H(S) = \sum_{x \in X} -p(x) log_2 p(x)$$

Where S is the current dataset for which entropy is being calculated, capitalized X is the set of classes in the label of the dataset, and $p(x)$ is the probability of each element in class x. Information gain is a concept to measure how much more information we could acquire from the dataset if we analyze the data one further step. The greater information gain is, the larger uncertainty or the more information we could obtain from the dataset if mining the data more. Information gain is usually denoted by $IG(A; S)$, where A is an attribute and S is the current dataset. It could be calculated from the following formula.

$$IG(A; S) = H(S) - H(A|S)$$

$$IG(A; S) = H(S) - \sum_{a \in A} \frac{S_a}{S} H(S_a)$$

where $H(S)$ is the entropy of dataset S, a is an element in attribute A, S_a is a collection of tuples whose A attribute value is a.

4 CONCLUSION

This paper introduces the Titanic story and how historians preprocess the data. The ID3 algorithm is also introduced for historians to employ on their research. As a historian, big data is an opportunity but also a big obstacle for the future researches. But at least, it is a good start in big data for all historians.

[2]

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

[4] Unknown. 2017. ID3 algorithm. Online. (10 2017). https://en.wikipedia.org/wiki/ID3_algorithm

Big Data Applications in Historical Studies

Neil Eliason
Indiana Unversity
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 Repertoire 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 it's 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 ProjectHistorical 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=edsoaf&AN=edsoaf.a29ec0ac934f1257030b477fa5986b1cff6def96&site=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=edsdoj&AN=edsdoj.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

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 INFRASTRUTURE

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 citys 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 the 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 worlds 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 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" [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 plight. The objective is that farmer 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-goverment 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 worlds 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 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 Develping World. Web page as Article. (Sept. 2014). 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

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.

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

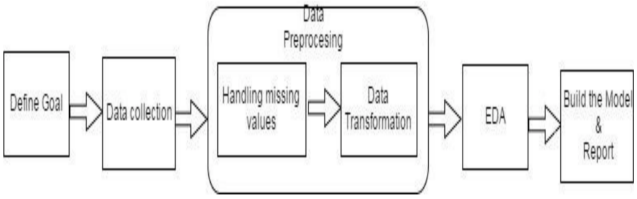


Figure 1: Steps in the predictive modeling process [3]

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.

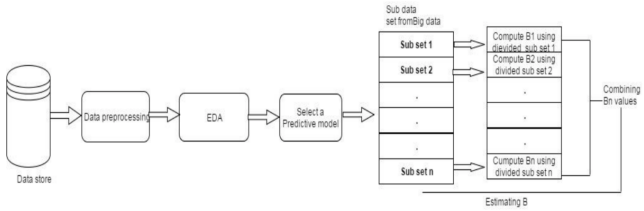


Figure 2: Architecture for partitioning Big Data [3]

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, $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_m$. Combine each of these parameters to compute the final regression coefficient, $\hat{\beta}_c$ as below:

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

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]. Some of the available

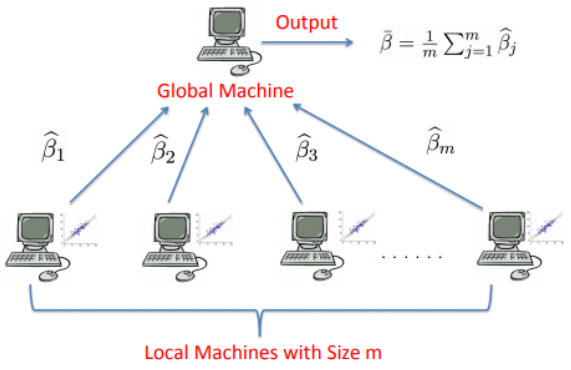


Figure 3: A divide-and-conquer learning framework [4]

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, β_0 gives the intercept and β_1 gives the slope of the regression line. Once the coefficients, β_0 and β_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] Sunghae 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

[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, Thiruvanthapuram, 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/

Big Data and Support Vector Machines

Dhawal Chaturvedi
Indiana University
2679 E. 7th St, Apt. C
Bloomington, Indiana 47408
dhchat@iu.edu

ABSTRACT

This paper provides an introduction to Support Vector Machines(SVM) and Big Data and tries to demonstrate how SVM can be used for classification of Big Data.

KEYWORDS

Big Data, Support Vector Machines, hid204, i523

1 INTRODUCTION

We live in a world increasingly driven by data. As the world is growing, the amount of data that is being created and stored on a global level is almost inconceivable, and it just keeps growing. As the amount of data increases, insights about it are relatively rare. It is getting increasingly difficult to get meaningful insights from these large datasets. Because this huge amount of data is unstructured aswell, it cannot be stored or processed in a traditional RDBMS or any other traditional Database Management System. Support Vector Machines(SVM) are supervised learning models that can be used to classify data according to their labels. We are discussing SVM because they are conceptually easier to understand when compared to other supervised learning algorithms. They can be applied to the same class of problems as other supervised learning algorithms along with neural networks. The only downside of using SVM is its complexity. Its gets even harder to implement an SVM for a huge dataset but if the training of an SVM can be broken down into smaller parts, then one can overcome this problem.
The paper is organised as follows. It gives an introduction to Support Vector Machines and talks about different types of SVM and ways to implement them. Then we describe different techniques such as hadoop framework through which we can implement SVM on Big data and gather meaningful information out of it.

2 SUPPORT VECTOR MACHINES

Support Vector Machine (SVM) was introduced by Vladimir N. Vapnik in 1995. It is one of the most popular learning algorithm which uses supervised learning for linear classification problems. They are based on the concept of decision boundaries which separates a set of objects having different features. Following are different types of SVMs. Along with linear classification, SVM is also capable of performing non-linear classification of dataset using the Kernel trick.

2.1 C-SVM classification

This kind of SVM is for binary classification usage. It is controlled with the parameter C which basically tells SVM optimization how much you want to avoid misclassifying each training example. The range of C can go from zero to infinity. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. In layman terms, C decides the width of your decision plane.

2.2 nu-SVM classification

This kind of SVM is also for binary classification usage. It is controlled with the parameter nu whose value can range from 0 to 1. This type of classification was introduced due to problems with C-SVM such as C can take any positive value and that it does not have any direct interpretation. In response Scholkopf et al. reformulated SVM to take a new regularization parameter nu which is bounded between 0 and 1 and has a direct interpretation that it is related to the ratio of support vectors and the ratio of the training error.

2.3 epsilon-SVR regression

In epsilon-SVR we do not have any control on how many data vectors from the dataset become support vectors, it could be a few, it could be many. But we will have total control of how much error to be allowed in our model to have, and anything beyond the specified epsilon will be penalized in proportion to C, which is the regularization parameter.

2.4 nu-SVM regression

In nu-SVR, the parameter nu is used to determine the proportion of the number of support vectors we desire to keep in our solution with respect to the total number of samples in the dataset. In nu-SVR the parameter epsilon is introduced into the optimization problem formulation and it is estimated automatically (optimally) for us.

2.5 Kernel trick for SVM

SVM can also be used for non-linear classification using the kernel trick. Kernels are functions which takes low dimensional input space and transform it to a higher dimensional space. In layman terms, what it does is, it converts data which is not linearly separable in lower dimension to a higher dimension in which it can be separated linearly.

3 PARALLEL SVM FOR BIG DATA

To fasten the process of training SVM, parallel methods have been proposed by splitting the traing data into smaller subsets and training a network to assign samples of different subsets.

3.1 Need for Parallel SVM

The biggest problem with SVM is its algorithmic complexity and high memory requirements, especially with large datasets. We can overcome this issue through parallel implementation of the algorithm so that it can work more efficiently.

3.2 Architecture

Parallel SVM can be implemented on large datasets by breaking the dataset into smaller fragments and use a number of SVMs to process each individual data set and finding local support vectors. By doing this the overall training time can be reduced significantly as the processing time is divided between individual nodes [2].
The training is done on Partial Support Vector Machines(PSVMs). Every PSVM gives an incomplete solution which is local to that PSVM. This partial solution is further used to find the final complete solution[2]. Through this model, large data optimization work can be distributed into several individual small optimizations. The resultant support vectors of the previous node is given as an input to the next node. The output set of support vectors of two Support Vector Machines are merged into single set and used as an input for the next Support Vector Machine. This process is continued untill only 1 set of Support Vector Machine is left.

4 MAPREDUCE BASED PARALLEL SVM

Map Reduce methodology can also be used to implement Parallel Support Vector Machines. The entire dataset is divided into n equal parts. The sub datasets is used as an input to the computational unit. MapReduceDriver starts the MapReduce job on each node. Map job is performed on each node on their respective datsets. The output of the mapper task, which is a trained support vector, is sent to the reducer to perform reduce operation. In the reduce task, the global weight vector is being computed by taking all local support vector computed at individual mapper nodes as an input. The training procedure will iterate until all sub-SVM are merged into one SVM [2].

5 QUANTUM SUPPORT VECTOR MACHINE

Not just high memory but the algorithmic complexity makes it impossible for SVM to be implemented on a large dataset. Patrick Rebentrost et al proposed that a quantum support vector machine can be implemented with algorithmic complexity of O(log NM) in both training and classification stages. It can be done due to a fast quantum evaluation of inner products and re-expressing the SVM as an approximate least-squares problem that allows for a quantum solution with the matrix inversion algorithm [3]. A technique for the exponentiation of non-sparse matrices which is recently developed allows us to reveal efficiently in quantum form the largest eigenvalues and corresponding eigenvectors of the training data overlap (kernel) and covariance matrices [3]. We can therefore, efficiently perform a low-rank approximation of these matrices, which is also known as Principal Component Analysis(PCA). In cases when a low-rank approximation is appropriate, this quantum SVM operates on the full training set in logarithmic runtime which is enormous improvement from O(n^2) [3].

6 PARTICLE SWARM OPTIMIZATION ALGORITHM

The particle swarm algorithm (Particle Swarm Optimization, PSO algorithm), is based on an idea of possibility to solve the optimization problems using modeling of animal groups' behavior. "PSO algorithm consists of the repeated applications of the fixed type of the kernel functions to choose optimal values of the kernel function parameters and value of the regularization parameter with the subsequent choice of the best type of the kernel function and values of the kernel function parameters and value of the regularization parameter corresponding to this kernel function type" [1]. The PSO algorithm and other swarm optimization algorithms are quite suited for the distributed architecture and handling of high volume unstructured data in the Big Data analytics because of the similarity between the behavior of swarms and big data, both being unpredictable.
Liliya Demidova et al. proposed a modified version of PSO algorithm which can be implemented for big datasets. The modified PSO algorithm conducts a simultaneous search of the type of kernel functions, the parameters of the kernel function and the value of the regularization parameter for the SVM classifier. "The idea of particles regeneration served as the basis for the modified PSO algorithm" [1]. In the implementation of this algorithm, some particles change the type of their kernel function to the one which corresponds to the particle with the best value of the classification accuracy. The PSO algorithm which is proposed reduces the time consumed by the developed SVM classifiers, which is an important factor for Big Data classification problem. In majority of the cases, these classifiers provide a high quality of data classification. In few exceptional cases the SVM ensembles based on the decorrelation maximization algorithm for the different strategies of the decision-making on the data classification and the majority vote rule can be used [1]. Along with this, a two-level SVM classifier has been proposed which works as the group of the SVM classifiers at the first level and as the SVM classifier on the base of the modified PSO algorithm at the second level. The promising results support the efficiency of these two approaches for classification of Big datasets [1].

7 CONCLUSION

Unlike the popular belief, Support Vector Machines can be used for Big Data analysis . People usually believe that they cannot be used for big data classification because of their highly complex nature. But as we have seen from the above examples, a lot of work has been done in this field to make SVMs more compatible with big data. The approach in most cases is trying to replicate the HDFS architecture by replacing each node with a partial SVM.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper. The author would also like to thank Mr. Aditya Tandon for proof reading his paper and giving suggestions to improve it.

REFERENCES

[1] L. A. Demidova, E. V. Nikulchev, and Yu. Sokolova. 2016. The SVM Classifier Based on the Modified Particle Swarm Optimization. *CoRR* abs/1603.08296 (2016). arXiv:1603.08296 http://arxiv.org/abs/1603.08296

[2] Anushree Priyadarshini and Sonali Agarwal. 2015. A Map Reduce based Support Vector Machine for Big Data Classification. 8 (10 2015), 77–98.

[3] P. Rebentrost, M. Mohseni, and S. Lloyd. 2014. Quantum Support Vector Machine for Big Data Classification. *Physical Review Letters* 113, 13, Article 130503 (Sept. 2014), 130503 pages. https://doi.org/10.1103/PhysRevLett.113.130503 arXiv:quant-ph/1307.0471

Clustering Algorithms in Big Data Analysis

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

ABSTRACT

Data Mining is a kind of popular method which intent to extract and analyze useful information from data. However, the rapid development of computing technologies has caused the size of data sets become extremely large. Because of the high complexity of these data sets, traditional data mining approaches or algorithms are not appropriate to be used. Therefore, it is very important to discover similarities of data and use them to divide data into different groups. Nowadays, clustering algorithms have emerged as a powerful meta-learning tool which provides the goal to categorize data into clusters such that objects are grouped in the same cluster are similar according to specific properties like traits or behaviors. In this paper, we have introduced several types of clustering technologies in data mining with some most commonly used algorithms, and compare advantages and disadvantages between them during the large data sets environment.

KEYWORDS

I523, HID209, Big Data, Clustering Algorithms, K-Means, CLIQUE

1 INTRODUCTION

Today, human's progress on Internet, Internet of Things (IoT) and other computing technologies lead to the growth of many related applications. Due to the usage of these applications in daily life, there are a huge amount of data generated every second which let the concept of big data emerged. Since these data may reveal hidden information which is interesting and important for people to study, the management of big data plays a significant role. We viewed technologies which used to extracting meaningful and required information or to find out the unseen relationship between the data as Data Mining [3], and clustering is one of the core tasks for processing data.

Clustering of big data has gained popularity in the last decade due to increased demand for a process of large data sets [2]. Data clustering algorithms were developed as a meta-learning tool to analyze massive data accurately. Its main purpose is to find similarities between all objects in a given data set and categorize them into groups by specific metrics which means that clustering is a process to manage similar objects in the same group. And since the data set processed by a clustering algorithm is unlabeled, data clustering has also considered as an unsupervised learning technique [1].

Because of the large amounts of data produced from different sources in today's world, the demand of efficiency clustering algorithms is increasing. However, it is difficult to create a perfect algorithm which has the capability to satisfy the requirement of all situations. Thus, depending on the purposes of clustering and the given data set, people should utilize the corresponding clustering approach [2]. Here are the main types of clustering (algorithms) which commonly used:

- Exclusive or partitioning-based clustering;
- Hierarchical clustering;
- Density-based clustering;
- Grid-based clustering;
- Model-based clustering.

In the rest of this paper we will discuss these cluster types as well as some of the applications.

2 CLUSTERING TYPES

Since each clustering approach has its own idea for solving the specific problems. This section introduces five different clustering types respectively with specific implementation procedures.

2.1 Exclusive or Partitioning-based Clustering

Partitioning is a kind of technique which used to decompose the set of data objects into several non-overlapping partitions (each partition represents a cluster) such that each data object is in exactly one partition.

Given a set of n data points, a partitioning method usually starts with a random partitioning and refine it iteratively to find a partition of k ($k \leq n$) clusters that optimizes the chosen partitioning criterion. While classifying data points into k groups, it must satisfy two rules: one point must be present in each cluster and one cluster must have one set [3]. One of the most famous and commonly used partitioning techniques is k-means.

2.2 Hierarchical-based Clustering

Hierarchical clustering groups a big scale of data set by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where closeness clusters at one level of leaf nodes are joined as new clusters that compose to nodes at the next level [4]. Generally, there are two approaches applied in hierarchical clustering which is the agglomerative method (top-down approach) and divisive method (bottom-up approach).

For the agglomerative method, it starts in two or more clusters recursively merged as the most applicable cluster by moving up the hierarchy. However, for the divisive method, it starts in a single cluster and recursively split to find applicable cluster for the items by moving down the hierarchy [3]. Although the time complexity for both approaches are huge ($O(n^2 \log(n))$ for agglomerative and $O(2^n)$ for divisive) which makes them execute too slow during the big data sets, it is possible to choose optimal agglomerative methods such as SLINK for single-linkage and CLINK for complete-linkage clustering to reduce the complexity to $O(n^2)$. BIRCH, CURE, ROCK

and Chameleon are some typical algorithms that applied under hierarchical-based clustering.

2.3 Density-based Clustering

Density-based clustering identifies distinctive clusters in the data based on the basic idea that a cluster in a data space is a contiguous region of high point density separated by regions of lower object density [5] and defined as a maximal set of density-connected points. It has the capability to learned clusters with arbitrary shape. Techniques such as DBSCAN, OPTICS, DBCLASD and DENCLUE are used to sifter out outliers and determine clusters of arbitrary shape [3].

$\varepsilon - Neighborhood$, which means objects within a radius of ε from an object, could define the density. It has the following form.

$$N_{\varepsilon}(p) : \{q | d(p,q) \leq \varepsilon\}$$

Where q is the neighborhood of cluster p. A "high density" neighborhood of an object contains at least MinPts (a specified number of points) of objects. If a point has more than MinPts objects within ε, it is a core point. If a point has fewer than MinPts objects within ε but in the neighborhood of a core point, it is a border point. A noise point is neither a core point nor a border point.

2.4 Grid-based Clustering

As the name of this clustering technique, it separates data objects in the data sets into different grids. Unlike other clustering methods that concerned with the data points, grid clustering focused on the value space that surrounds the data points. In general, a typical grid-based clustering algorithm has the following five steps:

1 Creating the grid structure. For example, partitioning the data space into a finite number of cells;
2 Calculating the cell density for each cell;
3 Sorting the cells by their densities;
4 Identifying centers of a cluster;
5 Traversal of neighbor cells.

Grid-based clustering has the similar time complexity to other clustering methods. However, it could significantly reduce the statistical value calculation complexity even though for clustering very big data sets. Some typical examples of this clustering are the Wave-Cluster and STING [3].

2.5 Model-based Clustering

Model-based clustering was developed based on probability models from the data. For example, the finite mixture model for probability densities. To find the parameter insider the probability model, it uses maximum likelihood estimation (MLE). In the model-based clustering approach, the data comes from a distribution that is a mixture of two or more components. Each component is described by a density function and has an associated probability in the mixture. Let assume the components adopt $p - variate$ normal distributions. Then the probability model for clustering will often be a mixture of multivariate normal distributions and each component in the mixture will represent a cluster. Nowadays, there are many good techniques that could be used in model-based clustering such as expectation-maximization (EM) algorithm, conceptual clustering and neural network approaches.

2.6 Comparison of Clustering Methods

As we have discussed some of the mainly used clustering methods, here we could assess them and sum up their advantages and disadvantages within the big data environment. Table 1 displays this information for each clustering type.

3 CLUSTERING ALGORITHMS

Here in this section introduces two kinds of well-known and commonly used clustering algorithms: the k-means algorithm which applied partitioning-based clustering and CLIQUE algorithm which applied grid-based clustering.

3.1 K-Means Algorithm

K-means algorithm is a powerful method to exploring the structure in a data set [6]. It classifies the whole data set with n instance into k clusters and aims to find an optimal solution that minimizes the value of objective function. K-means uses centroid which is the average points to represent a cluster. And then in assignment step, it obtains the distance between the data point and the centroid to find the nearest cluster for each point. After that, update centroid by recomputing them of each cluster. The procedure of k-means algorithm is simple which contains the following main steps:

1 Randomly select k points as initial centroids;
2 Calculate the distance between each data point and centroids;
3 Assign each data point to the centroid with the minimum distance;
4 Repeat the calculation on the centroid of each cluster until centroid does not change.

The time complexity of the k-means algorithm is $O(nk)$ which could be extremely expensive when both n and k are large. However, there are several ways to overcome this problem such as reduce the cluster number in assignment step or quickly identify data points that often change the cluster. By implementing k-means algorithm through the Map-Reduce framework, it will have the ability to handle big data [6].

3.2 CLIQUE

CLIQUE, also known as Clustering in QUEst, is used to find density regions from a sparse multi-dimensional data set [6]. Each region could be identified as a cluster through properties like attribute values, points or ranges. Since CLIQUE algorithm has the ability to discover density units automatically, it is able to be applied to higher dimensional data sets which are one of its advantages compared to the other clustering algorithms. CLIQUE partitions m-dimensional data space into a rectangular unit without overlapping in order to get the dense units. And clusters are generated from the subspaces of original data spaces through the Apriori property [6]. Assume there is a dense unit which is k dimension, its projections are $(k-1)$ dimension. Thus, by using CLIQUE algorithm, we could obtain the minimum descriptions for its data points. CLIQUE algorithm contains the following main steps:

1 Identify subspaces that contain clusters;
2 Identify clusters;

Clustering Type	Advantages	Disadvantages
Partitioning-based clustering	1 Relaxed to appreciate and implement; 2 Produce additional thick cluster than the hierarchical technique especially when clusters are circular; 3 For large number of variables, k-means algorithm may faster than hierarchical clustering when k is small; 4 Well-organized in processing large data sets.	1 Deprived at usage of noisy data and outliers; 2 Works only on numeric data; 3 Unfilled cluster generation problem; 4 Haphazard preliminary cluster center problem; 5 Not appropriate for non-spherical clusters; 6 User has to provide the value of k.
Hierarchical-based clustering	1 More adaptable; 2 Less delicate to noise and outliers; 3 Any amount of clusters can be acquired by cutting the dendrogram at desired level. It allows diverse users to choose dissimilar panels according to the desired resemblance level; 4 Appropriate to any characteristic type.	1 If a process is performed, it cannot be undone; 2 Incompetence to scale well.
Density-based clustering	1 Resilient to outliers; 2 Does not necessitate the amount of clusters; 3 Forms clusters of uninformed shapes; 4 Unresponsive to organization of data objects.	1 Inappropriate for high-dimensional data sets due to the expletive of dimensionality singularity; 2 Its quality be contingent upon the threshold set.
Grid-based clustering	1 Fast handling time; 2 Self-governing of the amount of data objects.	1 Be contingent only on the amount of cells in each dimension in the quantized space.
Model-based clustering	1 Vigorous to noisy data or outliers; 2 Fast handling speed; 3 It decides the amount of clusters to produce.	1 Multifarious in nature.

Table 1: Comparison between each clustering algorithm [3].

3 Generate the minimum description for the clusters which identified in step 2.

4 CONCLUSION

The aim of this paper is to demonstrate the different clustering algorithms used in big data which included partitioning-based clustering, hierarchical-based clustering, density-based clustering, grid-based clustering and model-based clustering, compare their merits and demerits, and extend the content with their applications. Each clustering method adopts a different idea and implementation procedures to processes the data set which characterized by different properties. Thus, for choosing a clustering algorithm, we need to consider about both data sets, time complexity and other constraint conditions. For introducing clustering applications, it gave two simple but widely used algorithm cases: k-means and CLIQUE algorithm. K-means is a kind of partitioning clustering which provides effective results on data sets with numeric attributes, but this could be affected by noise or outliers and its time complexity might be very large in big data set. CLIQUE is a kind of grid clustering, it can find dense regions from multi-dimensional data sets and appropriate to deal with large volume of data. However, it could only apply to numerical data set as well.

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] Charu C. Aggarwal and Chandan K. Reddy. 2014. Data Clustering Algorithms and Applications. *Taylor & Francis Group* (2014).
[2] Rasim Alguliyev, Ramiz Aliguliyev, Adil Bagirov, and Rafael Karimov. 2016. Batch Clustering Algorithm for Big Data Sets. *IEEE Conference* (October 2016).
[3] Dr. Meenu Dave and Remant Gianey. 2016. Different Clustering Algorithms for Big Data Analytics: A Review. *IEEE* (November 2016).
[4] MathWorks. 2017. Hierarchical Clustering. (2017). https://www.mathworks.com/help/stats/hierarchical-clustering.html?requestedDomain=www.mathworks.com
[5] Joerg Sander. 2010. Density-Based Clustering. *Springer* (2010), 270–273.
[6] Ajin V W and Lekshmy D Kumar. 2016. Big Data and Clustering Algorithms. *IEEE* (April 2016).

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.

$$\nabla = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

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.

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

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 shear 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 ∈ 1...n **do**
3: Pick coordinate $i \in 1....n$
4: $x_i^{t+1} = x_i - \lambda[\nabla^t f(x_i)]$

Here λ 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 ∈ 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 ∈ 1...n **do**
3: select $i = argmax(\nabla^t f(x_i))]$
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 ϵ
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 ($\{c^1, ..., c^m\}, T, n, w_o, k$)

1: **while** no convergence **do**
2: **for** machine ∈ $\{1....k\}$ in parallel **do**
3: $v_i = SGD(\{c^1, ..., c^k\}, T, n, w_o)$
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 θ 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 θ_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 ∇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 shear 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
[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

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 individualfis 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:

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_n \end{array} \begin{array}{c} \begin{array}{cccc} fbPage_1 & fbPage_2 & \dots & fbPage_n \end{array} \\ \begin{pmatrix} 1 & 0 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix} \end{array}$$

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:

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_n \end{array} \begin{array}{c} \begin{array}{cccc} user_1 & user_2 & \dots & user_n \end{array} \\ \begin{pmatrix} 1 & .75 & \dots & .85 \\ .75 & 1 & \dots & .91 \\ \vdots & \vdots & \ddots & \vdots \\ .85 & .91 & \dots & 1 \end{pmatrix} \end{array}$$

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 peoplefis 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 personfis 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 userfis 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 userfis 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 fisocial creditfi scheme plan. eNewsPaper. (July 2017). https://www.ft.com/content/f772a9ce-60c4-11e7-91a7-502f7ee26895 China changes tack on fisocial creditfi 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.

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 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 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].

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.

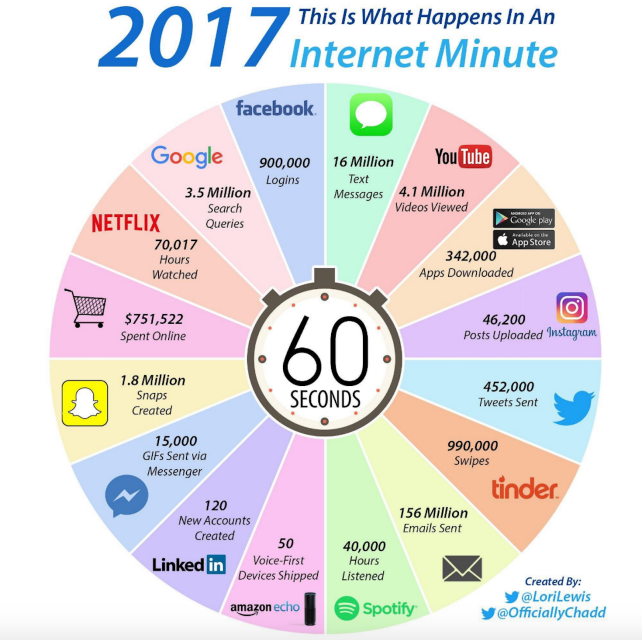


Figure 1: What happens in an internet minute? [8]

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.

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 to like to thank the associate instructors for being available on the course website all the time and helping us with their answers.

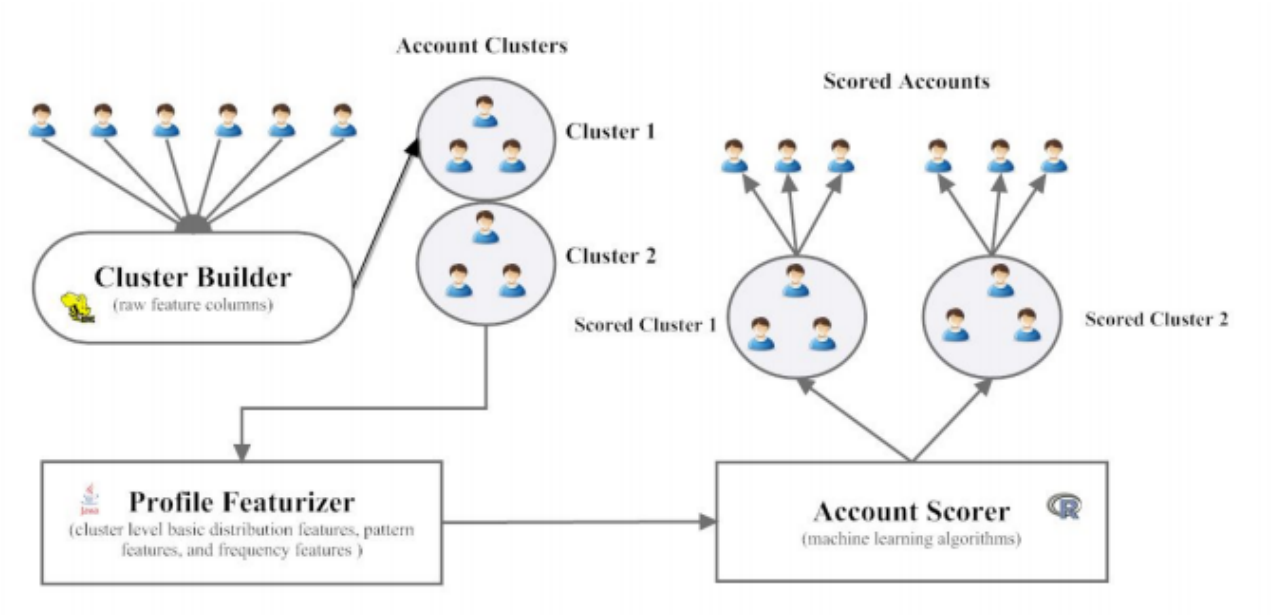


Figure 2: Fake User detection approach used by LinkedIn[21]

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 Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant 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

Netflix use of Big Data Visualization

Zachary Meier
Indiana University
Bloomington, Indiana 47408
zrmeier@indiana.edu

ABSTRACT

When you think of Netflix you probably don't think about big data. You think about streaming videos and being watching the newest series. How they keep you watching and a happy customer is through their use of big data. Not only does Netflix use big data for their external customers they use it for their internal customers. Netflix use of big data analytics has given them strong customer insight and internal applications to help keep their systems up and running and striving for the four nines availability. The uptime of 99.99% which they strive for and have tried to uphold very well.

KEYWORDS

Big Data, Edge Computing i523

1 INTRODUCTION

Netflix was once a small start up company, trying to compete with Blockbuster and Hollywood Video. It's main goal was to allow customers to get DVD's through the mail without having to go in store. They expanded their services through distributing an API to other companies to encourage streaming and coming back to their site. That originally turned out to be a failure, but then they started to see the value in streaming, and started to allocate major resources in that direction. From there they became a big data giant in the movie streaming business.

2 BRIEF HISTORY

There is no doubt you have not heard of Netflix, but their climb to where they are now has been a journey of failure and strife. Netflix once was a DVD only shipping business, but soon started experimenting with streaming by pushing out their API to other companies to drive business back to them. However this had been a huge failure, but through that they found value in streaming on their own. They started building teams and software to create to handle the data they encountered and originally started with their own infrastructure as well to host this service. Eventually they saw value in the use of micro services and the use of amazon web series as a solid company structure. Eventually they started living exclusively in the cloud. With that they faced many challenges but they knew that the pro's out weight the con's and pushed forward into the new frontier of cloud computing and micro service industry. Netflix now has almost 100 million views in almost every country around the globe. However, they plan to push for more customers. They plan to dedicate 1 billion dollars to acquiring new customers. It is uncertain what the future hold for Netflix but one can bet that it will involve big data, and that is what will drive them forward.[4]

3 THE REVOLUTION OF THE GENRE

If someone wanted to count how many genres were out there for movies, they would be able to count it easily. However, when Netflix was looking for a way to create better data so they could create better suggestions to retain customers, they revolutionized that system. Gone are the days of simplistic genres of the past created by Hollywood. Now are the days of 76,897 to categorize movies according to Netflix. [3] Soon this grew into much more.

3.1 The Netflix Quantum Theory

This was a document in which was created to come up with a method of tagging movies with genres. In this case they were called micro genres. This set out to distinguish movies from one another while making it easier for a generator to put together movies someone would most likely watch based on the genres provided. Such as quirky movie or romantic comedy featuring the rock. Does that actually exist? If it did, Netflix Quantum Theory Generator would produce something like that to get people to watch it. This is the basis for their predictive algorithm which they are always fine tuning to provide value to their customers.

4 INTERNAL USE OF BIG DATA

Their is no doubt of use for big data to help customers find the movies they would like to watch. However that is only one side of the issue. For Netflix to be effective they need to be able to take the data they get back from the customers and turn it into something tangible and usable, so they can keep creating value for their customers. This is included in their use of content creation, not just the utilization of the license agreement from the movie companies. They use all of this information to keep track of their ecosystem of micro services as well as see what users are doing and how to better serve them. Due to their overwhelming size and use of data, Netflix had to get creative and start thinking on better ways to collect and push this data across all their internal teams to produce a company that works in harmony.

4.1 Atlas

Atlas is essentially the monitoring tool that Netflix uses to make their operation work. They implemented this system back in 2013 and it has been running since. It allows for easy time series queries to beget back valuable data that scales. They did this by simplifying the structure of the data and the query structures. In addition most of this data was kept in memory so it as easily accessible so as not have to access a database and slow down the process. This allows the data to be pulled quickly from many different services that demand the data. [1]

4.2 Visualization

It is one thing to understand the data from a mathematical and theoretical level. However that only helps those who are easily associated with it. At Netflix, if it is not visualized then it doesn't serve a purpose. Which cover the three ways Netflix views data.

- Data should be accessible, easy to discover, and easy to process for everyone.
- Whether your dataset is large or small, being able to visualize it makes it easier to explain.
- The longer you take to find the data, the less valuable it becomes.[5]

With these view you can see how Netflix want to accurate rich accurate data. Not only that but make it better. Something a normal person can look at, and be able to make a informed decision. For instance, in the case of choosing a new shows title picture. They put out different version and see what gets chosen, and all the characteristics of why that particular one was chosen. Was it based off color, people in it, typography. They can usually those data points to help pick the best picture to have people watch that show. That is one of the key drivers for why their business has continued to flourish.

4.3 Micro services

Not only does all of this data help identify shows customer want to watch by informing the employee, but it also helps them run a better system. Netflix would be nothing without the data they are able to obtain for their ecosystems of micro services. The essential lifeblood of their operation platform. With the use of big data they can fin tune how they run their micro services and help maintain the four nines of uptime the hold themselves to. With this data, they spot what services are essential for the product to still run, though it may be limited. They also use a lot for testing suits for different types of failures in production under load. These then allow them to get data back on how the system performed and how to fix it without actually breaking anything at all.

5 CONCLUSION

In conclusion, this is only a small view of what makes Netflix a Big Data giant. They have faced many other issues in the past and over come them with agility. Normally if a company grew this fast it would start to get really messy and make some possibly bad decisions based of their data. Just ask UBER about that type of situation. However, that issue may have been to lack of leadership and planning more than anything. In any case Netflix has adapted and conquered and now is a model for many other platforms out there such as amazon video and apple TV. Whether or not the will get to the level of Netflix is possible, but the real question is, by the time they get to that level, where will Netflix be?[2]

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski and the TA's for their help and support.

REFERENCES

[1] Roy Rapoport Brian Harrington. 2014. Introducing Atlas: Netflixfis Primary Telemetry Platform. (December 2014). https://medium.com/netflix-techblog/introducing-atlas-netflixs-primary-telemetry-platform-bd31f4d8ed9a

[2] Josh Evans. 2014. Embracing Failure: Fault Injection and Service Resilience at Netflix. (June 2014). https://www.youtube.com/watch?v=9R710ry-Cbo

[3] Alexis C. Madrigal. 2014. How Netflix Reverse Engineered Hollywood. (January 2014). https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

[4] Bernard Marr. 2015. Big Data: How Netflix Uses It to Drive Business Success. (August 2015). https://www.smartdatacollective.com/big-data-how-netflix-uses-it-drive-business-success/

[5] Phil Simon. 2014. Big Data Lessons From Netflix. (March 2014). https://www.wired.com/insights/2014/03/big-data-lessons-netflix/

Big Data Security and Privacy

Andres Castro Benavides
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 43017-6221
acastrob@iu.edu

Uma M Kugan
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 43017-6221
umakugan@iu.edu

ABSTRACT

In recent history, the explosion of smart devices, social media platforms and the Internet of Things has resulted in major replication of data. This data is transferred at high speeds, in large volumes and in a variety of different platforms. Data is one of the biggest assets of companies and with the exponential growth of data comes an increasing problem: Security and Privacy. Security and privacy have become a major initiative for every organization that has or utilizes databases. These organizations need to protect their brands, avoid penalties, and, in worst case scenarios, avoid circumstances in which they may either lose a significant amount of business or their business as a whole. The computer science industry, alongside of these organizations, need to continue to develop ways to protect, utilize, and gain real-time insight from the data that each organization has. This paper is going to highlight security and privacy in Big Data, alongside of specific issues and challenges related to security.

KEYWORDS

Security and Privacy, Big Data Security, Big Data, i523, HID305, HID323

1 INTRODUCTION

Each organization has unique needs when it comes to Big Data. These needs cannot be described with one defined structure alone, and likewise, the information that they use does not come with defined data types. Because of this, there is the need for the Big Data Platform. Big Data is gaining more popularity because of its ability to connect to a number of devices in the so-called "Internet of Things"(IoT), producing a huge dump of data that needs to be transformed into information assets. It is also very popular to buy additional on-demand computing power and storage from public and private clouds to perform intensive data-parallel processing. These things not only create the way for Big Data expansion but also boosts security and privacy issues. Big Data security is the process of securing data and their processes both within and outside the organization. Big Data deployments are valuable targets for intruders and, because of this, security becomes a never ending concern for any organization. A single unauthorized user gaining access to an organization's big data could in and of itself acquire all the valuable information that the company possesses which could result not only in monetary loss but also be detrimental to its business and to its brand name. In current trends, security teams work towards continuously monitoring networks, hosts and application behavior across their organization's data. Traditional methods of securing firewalls are no longer enough to secure a company's data assets and Big Data platforms need to be secured with a mix of both traditional and newly developed security tools, as well as big data analytics for monitoring security throughout the life of the platform [11].

2 WHAT IS BIG DATA

Big Data, by definition of its name, is an extensive variety and heavy volume of data that can be entered or transferred at high velocity, and include data sets coming from dynamic sources of data and applies technologies to analyze these data sets. It is a term usually used to define huge and complex data sets that do not fit into any traditional system. Most recently, the term "Big Data" tends to refer to the use of predictive, user behavior analytics, or certain other advanced data analytics that extract value from data sets. These analytics provide more insights about the data which indeed help businesses understand their trends which will eventually, in good theory, help their growth. [8].

For example, a company that works with waste management, can collect data on the waste production and human activities from very diverse sources, then interpret the findings of Big Data to make optimal decisions [24].

3 BIG DATA NEEDS BIG SECURITY

The amount of data collected by organizations and individuals around the world is growing on a daily basis, and the volume of the data being collected is expected to continue to grow exponentially. It is believed that the 90% of the data we have currently have in the world has been collected in the past few years. Velocity, volume and variety of Big Data comes results in privacy, security and compliance issues as well. Some of the data stored in Big Data platforms is very sensitive and regulations need be put in place, strictly controlling specific aspects of the data and who has access to the data. Proper measures have to be taken to control any weaknesses to cyber threats.

There are requirements for security measures already in place. Big Data platforms are subject to compliance mandates by government and industry regulations, including GDPR, PCI, Sarbanes-Oxley (SOX), and HIPAA [13]. These measures place regulations on company practices and implementations that ensure proper data security and monitoring. These regulations are mandatory, and failing to comply could result in severe penalties, from heavy fines to legal actions.

While these requirements are important, traditional security mechanisms that have been in place for securing structured static data are no longer sufficient. With technological advances also comes a need to continually assess weaknesses in the new systems, to protect itself from new cyber threats and hacking strategies, and to create user friendly platforms for client that do not compromise

the data being collected or stored. These developments are often far ahead of regulation, and individual entities need to be continually monitoring and enhancing their platforms to ensure protection of its data and systems. Big Data needs bigger security to protect its data, applications and infrastructure. Securing data not only protects the brand, reduces costs and avoids any legal issues, it also helps in retaining the brand name and increases revenue and growth [16].

4 BIG DATA SECURITY CHALLENGES

Recent adoption of cloud storage has increased the amount of data collected by organizations and hence it has become of vital importance to secure these data platforms as well. Data security issues are generally caused by the lack of proper tools and measures provided by traditional anti-virus software. Routine security checks to detect patches are no longer enough to handle real time influxes of data. Streaming real time data demands a great amount of attention focused on security and privacy solutions. Databases are no longer static. Big Data security's motto is to restrict unauthorized users and intruders from getting into a platform and also to block the encryption of data both in-transit and at-rest. The adoption of cloud storage creates a need to pay particular attention to the in-transit, or the continually expanding and modifying databases. Big Data security tools must be in place at all stages of data i.e. on incoming data, data stored in the platform and also on the data that goes out to other applications or outside party [20].

4.1 Access Control

Access control, in the context of Big Data, is controlling who can access data by using security settings. The different platforms that use Big Data need to be able to identify critical data, data origination and also who has access to the data. In this capacity, data access is not only protecting from external access, but also protecting data from those who have internal access as well [15].

User access should be controlled via a policy-based approach that automates access based on user and role-based settings. This manages different level of approvals in order to regulate who has access to the critical data and to protect the big data platform against inside attacks [4].

4.2 Audit Control

Big Data analytics can be used to analyze different types of logs in order to identify malicious activity. It also can regularly audit all the working directories inside the organization in order to check for any unauthorized access to any sensitive or privacy data. In reality, not all attacks are identified in the exact moment when the attack occurs. In order to perform a root cause analysis of the incident, data security professionals need to have access to audit logs which allow them to trace attacks back to the point of entry, exact time, modifications or weaknesses. In case of data breach, some firms are required to turn over their audit logs to stakeholders and possibly affected companies and heavy fines are imposed for failure to comply [16].

4.3 Real Time Compliance Control

Real time security monitoring is always very challenging due to the number of false positive alerts generated by security programs. Because of the frequency of false positive alerts, they are usually ignored. Big Data analytics may help provide more meaningful insights that could result in real time detection .

4.4 Non Relational Databases Privacy

Non Relational Databases are still not fully matured. This poses a severe threat to securing the data and it is often difficult for security and governance team to keep up with the demand. NoSQL databases primarily focus on how to handle high volume of data without paying much attention to their security needs.

4.5 End-Point Input Validation

Many organizations collect their data from End-Point devices. It is very important to ensure that data coming from these devices is not infected. Proper steps must be taken to make sure data is coming from an authentic source and it is legitimate. Incoming data from End-Point devices such as smart phones is growing tremendously and filtering or validating data from these sources is a very big challenge [16].

4.6 Securing Transaction Logs and Data

Data in any organization many be stored at various levels (tiers) of the storage structure depending on the need and usage of the data. Increase in the transfer of data within the organization enforce for the need of auto-tiering for Big Data storage whereas auto-tiering does not maintain the log of where the data is stored and hence security is a big concern.

4.7 Securing Distributed Framework

Distributed framework enforces parallelism. This means that data is distributed across multiple nodes to achieve faster processing of large volumes of data. This increases the security concern of the framework and the data that exists there. Most companies use a distributed framework like MapReduce in which mappers read and compute and reducers combine the output from each mapper. If mappers are not secured, there is the chance of data being compromised [4].

4.8 Data Provenance

It is very important to know the original data that is coming to the platform so that we can better classify them. Data Origin should be consistently monitored but in reality due to the high volume it is becoming a big concern for data security. Provenance metadata is growing significantly as well and protecting metadata is very crucial for any organization [4].

5 BIG DATA SECURITY STAKEHOLDERS

In the digital era, the traditional way of securing the data, changing passwords frequently, firewall protection is just not enough to keep up with the growth of data produced by Internet of Things(”IoT), Smart Devices, Bring Your Own Devices (”BYOD”) and several customer friendly apps that is coming out everyday. ”Even though

end user has the biggest responsibility with securing his own data, unfortunately, end users are not fully aware of the cyber security issues and they do not have the appropriate knowledge to discover the world wide web in complete safety" [17].

Big Data deployment is not possible to handle by any single business unit or with single tech team. It involves several business units, infrastructure, information technology, security,compliance, programmers, testers and product owners are all involved in big data deployment. They are all responsible for Big Data Security. Information Technology and Security team is responsible for drawing the policies and procedures. Compliance officers together with security team will protect compliance, such as automatically encrypting personally identifiable information before it is easily accessible. Administrators will automate these process to protect their environment. Even though every organizations have their policies and control laid in place to protect their biggest asset, phishing attacks can come in any form as a simple email. Frequent internal audit within the company can help us periodically check if all privacy, security and compliance are all in place.If not, proper measures can be taken right away to avoid any legal issues.

"The average annualized cost of cyber crime based upon a representative sample of 237 organizations in six countries by Ponemon Institute in their 2016 Cost of Cyber Crime Study and the Risk of Business Innovation sponsored by Hewlett Packard Enterprise is 9.5million U.S. dollars" [6]. In any organization, loss of information is the most expensive consequence of a cyber crime. The cyber attack may results in business disruptions, data or information loss, loss of revenue, damage to equipment and last but not the least it damages the brand. So it is big time to protect and secure the big data and the environment from all angles.

6 BEST PRACTICES FOR SECURING BIG DATA

There are three fundamental principles used in defining security goals: confidentiality, integrity, and availability. Confidentiality is the ability to keep sensitive data safe from third parties and unauthorized access. Integrity in this context means to avoid unauthorized modification of the data. Finally, availability means always being able to access the data and resources. These three concepts are known as the CIA triad, and is used as base principles when discussing and designing security practices [7].

To meet these goals, there are four main branches of security that apply to Big Data: Authentication, Encryption, Data Masking and Access Control [1].

6.1 Authentication

Because of its nature (large sizes of data, linking different sources, sharing access with third parties, etc), some of Big Data's features are highly susceptible to different privacy, security and welfare risks [9].

Privacy can be defined as the condition of confidentiality, protecting information from third parties. To support privacy, there have been different Authentication methods that both verify and validate entities who attempt to access the information. This ensures that only authorized entities are able to access the data or resources.

With Big Data, it is important to choose a proper authentication method, with the least computation complexity as possible, to allow dynamic security solutions within large Data Centers and also to avoid incrementing the traffic unnecessarily. Choosing an overbearing authentication method can cause both delay and storage issues. Because of this, it is important to tailor the security to the needs of the specific network [22].

6.2 Cryptography

There are multiple understandings of how data moves through stages, also known as Data Life cycles. Cryptography- define in terms of security. CITA.

From the perspective of cryptography, there are three phases in the Data Life Cycle: Data in Transit, Data in Storage, and Data in Use. Different cryptography techniques will be implemented depending on which stage of the life cycle the data is in [7].

There are different cryptographic tools that not only keep data secure at each point in its life-cycle, but also enable richer use of the data. The main tool is Encryption. Encryption takes pieces of data in plain text and use a cryptographic key to produce a version of the data that can only be read using the cryptographic key. Without the key, the information is illegible. There are two types of encryption: secret key encryption and public key encryption. Secret key encryption is when the same key is used for both encrypting and decrypting data. There are scenarios when one of the keys can be made public. For instance, if the locking key is kept private but the unlocking one is made public, this security can be used to prove authenticity [7].

There are different standards for encryption. The most well known and commonly used is Advanced Encryption Standard (AES). This standard sets guiding principles to ensure that data is encrypted in a manner that meets security needs and allows the recovery of original data [7].

6.3 Data Masking

By definition, Big Data works with large volumes of heterogeneous data sets using software to manage the data and to provide predictive analysis. Data masking works by replacing sensitive data with non-sensitive values, yet preserves the data integrity. For instance, replacing names with code names, or social security numbers with a key number. By doing this, different parties can access information without putting sensitive data at risk [2].

Five laws for data masking have been developed by Securosis Research. The first law is that data masking should not be reversible. This means that the data should not be unmasked easily using reverse engineering. The second law is that data that has been masked has to represent the original data set. For example, it has to belong in the same context. The third law states that data masking should maintain application and database integrity. This means that the process of data masking should not modify or affect the data in the databases in a negative way. The fourth law emphasizes that non-sensitive data can be masked, but it should not be masked if it can not be used to make sensitive data vulnerable. For instance, when masking information about a person, it is correct to mask the personfis name, email address and social security number, but other information like gender, or favorite colour, would be useless

to mask. Finally, data masking must be a repeatable process, using a standard to reproduce the steps taken to mask the data, allowing to troubleshoot possible problems in the process [2].

6.4 Access Control

As it was explained in the Challenges section, Access control, allows some entities to access the data or resources, while denying its use to other entities. Through security settings.

Some authors add that the inferences drawn from data should also be a cause for concern, because they can identify traits and patterns that could expose vulnerabilities. They propose that organizations who use the protected data should disclose their decisions criteria in order to apply access control in a broader spectrum. By doing so, it would be sufficient to diminish privacy concerns by de-identifying the data, or denying access to certain parts of the data that coul be used to make entities or data vulnerable. Some of these authors say that by doing this, it would not only reduce the privacy risk, it would also salvage large amounts of data for alternative use. This de-identification can also be achieved through data masking, pseudonymization, aggregation, among other methods [21].

6.5 Physical Security

It is always better to build and deploy Big Data platforms in their own data center. If deployed in a cloud, the organization must diligent to ensure that the cloud provider's data center is physically well secured. Access should be restricted to strangers and staff who have no official responsibilities in the designated areas or interacting with the data sources. Data centers should be properly monitored at all times and video surveillance and security logs are important tools to achieve this.

7 FUTURE OF BIG DATA SECURITY

To think about Future of Big Data Security, it is necessary to engage the conversation of what the trends are in Big Data and what technologies are expanding and changing the horizon. There are many new technologies and solutions that are shaping the future of the Big Data, but because of the length and focus of this document, there are three main areas that will be covered: Virtualization and Cloud Computing, IOT Security, External Password Vaults and Penetration Tests.

7.1 Virtualization and Cloud Computing

Virtualization is a way of deploying resources at multiple levels, such as hardware, network infrastructure, application and desktop centralized managing and using dynamically the physical resources. This makes the system flexible and less costly than traditional environments and giving management new tools to optimize the use of resources [14].

Since virtualization can be developed in so many levels, including cloud computing and by multiple service providers, it is natural that the system requirements of users and organizations move towards a variety of solutions that may include Infrastructure as a Service (IaaS) frameworks from public clouds such as the ones offered by Amazon, Microsoft, Google, Rackspace, HP, among others, or even Private clouds, maintained and many times even set up by internal IT departments [23].

These cloud computing technologies are being used to solve data-intensive problems on large-scale infrastructure. Thus, integrating big data technologies and cloud computing for data mining, knowledge discovery, and decision-making [10].

7.2 IOT Security

The Internet of Things (IoT) is the name given to the large network of physical devices that does not match the typical concept of computer networks, this includes all kinds of objects. The large and growing amount of devices and diverse uses given to them, makes IoT generates very important Big Data streams. Making it necessary to develop new systems and data mining techniques for this new paradigm [3].

In this IoT paradigm, each new opportunity opens doors to new threats as well. This makes it necessary to develop techniques to ensure trust, security and privacy. Different Authors write about the possible ways to face these challenges, and some, they consider three main axes to articulate the solutions: Effective security - used in very small embedded networks,context-aware privacy and user-centric privacy, and the third one is the systemic and cognitive approach for IoT security - where the interaction between people and the IoT can be envisioned as a set of nodes and tensions [18].

All this to say that in order to approach privacy and security in this new paradigm, many new theories and techniques have been developed since old security products and techniques may not suffice the needs of the different IoT users and communications.

7.3 External Password Vaults

Password vaults are applications that store multiple passwords and encrypt them storing them in a database [5].

There are small Password Vaults that can be stored locally on a system, or larger options that can be integrated into larger systems, providing additional security options, like generating real time temporary passwords for effective password rotation (I.E. Cyberark External Password Vault) [12].

These techniques are key to articulate authentication and a proper data access while using multiple services such as Cloud infrastructure and IoT.

7.4 Penetration Tests

After applying all the security techniques and strategies, and after putting in place all necessary security and privacy policies, the most important step is validating the strength of the security of the system. For some time, companies have started to perform tests that consist on simulating an attack from the perspective of an attacker, this method is known as Penetration test and it allows to actively evaluate and assess the security of a system [19].

The tester identifies the threats faced by an organization from hackers and suggest changes to improve the security and minimize the vulnerabilities and close the possible loop holes in the network [19].

8 CONCLUSIONS

Big Data as a constantly evolving and ever changing branch of information technologies resembles an ecosystem that since it covers gathering data from so many sources, processing it and generating

new information, there will be many entities and interests involved that will need to be protected.The features of Big Data such as Volume, Variety and Velocity bring new challenges to security and privacy protection. To protect the integrity and availability, security providers and local IT departments, will have to diversify their security and privacy strategies and policies, in order to keep pace with the growth and evolution of this new ecosystem.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Karim Abouelmehdi, Abderrahim Beni-Hssane, Hayat Khaloufi, and Mostafa Saadi. 2017. "Big data security and privacy in healthcare: A Review". *"Procedia Computer Science"* 113, Supplement C (2017), 73 – 80. https://doi.org/10.1016/j.procs.2017.08.292 The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.

[2] RA Archana, Ravindra S Hegadi, and TN Manjunath. 2017. A Big Data Security using Data Masking Methods. *Indonesian Journal of Electrical Engineering and Computer Science* 7, 2 (2017), 449–456.

[3] Albert Bifet. 2016. Mining Internet of Things (IoT) Big Data Streams.. In *SIMBig*. CEUR-WS.org, CUSCO, PERU, 15–16.

[4] Peter Buttler. 2017. *Big Data Needs Big Security. Herefis Why.* Technical Report. Dataconomy, Berlin, Germany. http://dataconomy.com/2017/07/10-challenges-big-data-security-privacy/

[5] Rahul Chatterjee, Joseph Bonneau, Ari Juels, and Thomas Ristenpart. 2015. Cracking-resistant password vaults using natural language encoders. In *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, CEUR-WS.org, CUSCO, PERU, 481–498.

[6] Ponemon Research Department. 2016. Ponemon Institute Research Report. (October 2016). https://www.ponemon.org/local/upload/file/2016%20HPE%20CCC%20GLOBAL%20REPORT%20FINAL%203.pdf

[7] Ariel Hamlin, Nabil Schear, Emily Shen, Mayank Varia, Sophia Yakoubov, and Arkady Yerukhimovich. 2016. *Cryptography for Big Data Security*. Taylor & Francis CRC Press, Boca Raton, Florida. 241–288 pages.

[8] A.J.G. Hey, S. Tansley, and K.M. Tolle. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, REDMOND, WASHINGTON. https://books.google.com.my/books?id=oGs_AQAAIAAJ

[9] Nir Kshetri. 2014. Big data s impact on privacy, security and consumer welfare. *Telecommunications Policy* 38, 11 (2014), 1134–1145.

[10] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya. 2016. The anatomy of big data computing. *Software: Practice and Experience* 46, 1 (2016), 79–105.

[11] Jose Moura and Carlos Serrao. 2016. Security and Privacy Issues of Big Data. *CoRR* abs/1601.06206 (2016), 1–2. https://doi.org/10.4018/978-1-4666-8505-5.ch002 arXiv:1601.06206

[12] Gregory S Nelson. 2015. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. In *SAS Global Forum Proceedings*. SAS Institute Inc., DALLAS, TEXAS, XXIII.

[13] Cheryl OfiNeill. 2017. *Big Data Needs Big Security. Herefis Why.* Technical Report. Imperva, Redwood Shores,California. https://www.imperva.com/blog/2017/02/big-data-needs-big-security-heres/

[14] K Padmini. 2015. Securing data management based on key technologies in cloud computing. *International Journal of Advance Research in Computer Science and Management Studies* 3, 2 (February 2015), 165–172.

[15] Amine RAHMANI, Abdelmalek AMINE, and Mohamed Reda HAMOU. 2015. A Mathematical MODEL OF ACCESS CONTROL IN BIG DATA USING CONFIDENCE INTERVAL AND DIGITAL SIGNATURE. *Computer Science & Information Technology* 5 (November 2015), 183–198.

[16] Sreeranga Rajan, Wilco van Ginkel, and Neel Sundaresan. 2012. *Top Ten Big Data Security and Privacy Challenges*. Technical Report. Cloud Security Alliance. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf

[17] Realdolmen. 2017. Cyber Security. (July 2017). http://www.realdolmen.com/en/blog/who-responsible-for-data-security-your-company

[18] Arbia Riahi, Enrico Natalizio, Yacine Challal, Nathalie Mitton, and Antonio Iera. 2014. A systemic and cognitive approach for IoT security. In *Computing, Networking and Communications (ICNC), 2014 International Conference on*. IEEE, ICNC, HONOLULU, USA., 183–188.

[19] Chaitra N Shivayogimath. 2014. AN OVERVIEW OF NETWORK PENETRATION TESTING. *International Journal of Research Engineering and Technology* 03, 07 (2014), 408–413.

[20] Christine Taylor. 2017. *Big Data Security*. Technical Report. Datamation, Foster City, California. https://www.datamation.com/big-data/big-data-security.html

[21] Omer Tene and Jules Polonetsky. 2012. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11 (2012), xxvii.

[22] Vijey Thayananthan and Aiiad Albeshri. 2015. Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. *Procedia Computer Science* 50 (2015), 149–156.

[23] Gregor von Laszewski, Fugang Wang, Hyungro Lee, Heng Chen, and Geoffrey C. Fox. 2014. Accessing Multiple Clouds with Cloudmesh. In *Proceedings of the 2014 ACM International Workshop on Software-defined Ecosystems (BigSystem '14)*. ACM, New York, NY, USA, 21–28. https://doi.org/10.1145/2609441.2609638

[24] Vitthal Yenkar and Mahip Bartere. 2014. Review on fiData Mining with Big Datafi. *International Journal of Computer Science and Mobile Computing* 3, 4 (2014), 97–102.

A WORK BREAKDOWN

Research for Section Big Data Needs Big Security, Big Data Security and Challenges: Uma Kugan.

Research for Section Best practices and Future: Andres Castro Benavides.

Editing: Andres Castro Benavides and Uma Kugan.

Big Data Analytics and IoT Smart Refridgerators

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].

image missing

Figure 1: 2016 Gartner Hype-Cycle Chart.

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 the 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 centraliztion 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=iewDAAAAIBAJ&pg=3607,488766&hl=en

Big Data Security and Privacy

Andres Castro Benavides
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 43017-6221
acastrob@iu.edu

Uma M Kugan
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 43017-6221
umakugan@iu.edu

ABSTRACT

In recent history, the explosion of smart devices, social media platforms and the Internet of Things has resulted in major replication of data. This data is transferred at high speeds, in large volumes and in a variety of different platforms. Data is one of the biggest assets of companies and with the exponential growth of data comes an increasing problem: Security and Privacy. Security and privacy have become a major initiative for every organization that has or utilizes databases. These organizations need to protect their brands, avoid penalties, and, in worst case scenarios, avoid circumstances in which they may either lose a significant amount of business or their business as a whole. The computer science industry, alongside of these organizations, need to continue to develop ways to protect, utilize, and gain real-time insight from the data that each organization has. This paper is going to highlight security and privacy in Big Data, alongside of specific issues and challenges related to security.

KEYWORDS

Security and Privacy, Big Data Security, Big Data, i523, HID305, HID323

1 INTRODUCTION

Each organization has unique needs when it comes to Big Data. These needs cannot be described with one defined structure alone, and likewise, the information that they use does not come with defined data types. Because of this, there is the need for the Big Data Platform. Big Data is gaining more popularity because of its ability to connect to a number of devices in the so-called "Internet of Things"(IoT), producing a huge dump of data that needs to be transformed into information assets. It is also very popular to buy additional on-demand computing power and storage from public and private clouds to perform intensive data-parallel processing. These things not only create the way for Big Data expansion but also boosts security and privacy issues. Big Data security is the process of securing data and their processes both within and outside the organization. Big Data deployments are valuable targets for intruders and, because of this, security becomes a never ending concern for any organization. A single unauthorized user gaining access to an organization's big data could in and of itself acquire all the valuable information that the company possesses which could result not only in monetary loss but also be detrimental to its business and to its brand name. In current trends, security teams work towards continuously monitoring networks, hosts and application behavior across their organization's data. Traditional methods of securing firewalls are no longer enough to secure a company's data assets and Big Data platforms need to be secured with a mix of both traditional and newly developed security tools, as well as big data analytics for monitoring security throughout the life of the platform [11].

2 WHAT IS BIG DATA

Big Data, by definition of its name, is an extensive variety and heavy volume of data that can be entered or transferred at high velocity, and include data sets coming from dynamic sources of data and applies technologies to analyze these data sets. It is a term usually used to define huge and complex data sets that do not fit into any traditional system. Most recently, the term "Big Data" tends to refer to the use of predictive, user behavior analytics, or certain other advanced data analytics that extract value from data sets. These analytics provide more insights about the data which indeed help businesses understand their trends which will eventually, in good theory, help their growth. [8].

For example, a company that works with waste management, can collect data on the waste production and human activities from very diverse sources, then interpret the findings of Big Data to make optimal decisions [24].

3 BIG DATA NEEDS BIG SECURITY

The amount of data collected by organizations and individuals around the world is growing on a daily basis, and the volume of the data being collected is expected to continue to grow exponentially. It is believed that the 90% of the data we have currently have in the world has been collected in the past few years. Velocity, volume and variety of Big Data comes results in privacy, security and compliance issues as well. Some of the data stored in Big Data platforms is very sensitive and regulations need be put in place, strictly controlling specific aspects of the data and who has access to the data. Proper measures have to be taken to control any weaknesses to cyber threats.

There are requirements for security measures already in place. Big Data platforms are subject to compliance mandates by government and industry regulations, including GDPR, PCI, Sarbanes-Oxley (SOX), and HIPAA [13]. These measures place regulations on company practices and implementations that ensure proper data security and monitoring. These regulations are mandatory, and failing to comply could result in severe penalties, from heavy fines to legal actions.

While these requirements are important, traditional security mechanisms that have been in place for securing structured static data are no longer sufficient. With technological advances also comes a need to continually assess weaknesses in the new systems, to protect itself from new cyber threats and hacking strategies, and to create user friendly platforms for client that do not compromise

the data being collected or stored. These developments are often far ahead of regulation, and individual entities need to be continually monitoring and enhancing their platforms to ensure protection of its data and systems. Big Data needs bigger security to protect its data, applications and infrastructure. Securing data not only protects the brand, reduces costs and avoids any legal issues, it also helps in retaining the brand name and increases revenue and growth [16].

4 BIG DATA SECURITY CHALLENGES

Recent adoption of cloud storage has increased the amount of data collected by organizations and hence it has become of vital importance to secure these data platforms as well. Data security issues are generally caused by the lack of proper tools and measures provided by traditional anti-virus software. Routine security checks to detect patches are no longer enough to handle real time influxes of data. Streaming real time data demands a great amount of attention focused on security and privacy solutions. Databases are no longer static. Big Data security's motto is to restrict unauthorized users and intruders from getting into a platform and also to block the encryption of data both in-transit and at-rest. The adoption of cloud storage creates a need to pay particular attention to the in-transit, or the continually expanding and modifying databases. Big Data security tools must be in place at all stages of data i.e. on incoming data, data stored in the platform and also on the data that goes out to other applications or outside party [20].

4.1 Access Control

Access control, in the context of Big Data, is controlling who can access data by using security settings. The different platforms that use Big Data need to be able to identify critical data, data origination and also who has access to the data. In this capacity, data access is not only protecting from external access, but also protecting data from those who have internal access as well [15].

User access should be controlled via a policy-based approach that automates access based on user and role-based settings. This manages different level of approvals in order to regulate who has access to the critical data and to protect the big data platform against inside attacks [4].

4.2 Audit Control

Big Data analytics can be used to analyze different types of logs in order to identify malicious activity. It also can regularly audit all the working directories inside the organization in order to check for any unauthorized access to any sensitive or privacy data. In reality, not all attacks are identified in the exact moment when the attack occurs. In order to perform a root cause analysis of the incident, data security professionals need to have access to audit logs which allow them to trace attacks back to the point of entry, exact time, modifications or weaknesses. In case of data breach, some firms are required to turn over their audit logs to stakeholders and possibly affected companies and heavy fines are imposed for failure to comply [16].

4.3 Real Time Compliance Control

Real time security monitoring is always very challenging due to the number of false positive alerts generated by security programs. Because of the frequency of false positive alerts, they are usually ignored. Big Data analytics may help provide more meaningful insights that could result in real time detection .

4.4 Non Relational Databases Privacy

Non Relational Databases are still not fully matured. This poses a severe threat to securing the data and it is often difficult for security and governance team to keep up with the demand. NoSQL databases primarily focus on how to handle high volume of data without paying much attention to their security needs.

4.5 End-Point Input Validation

Many organizations collect their data from End-Point devices. It is very important to ensure that data coming from these devices is not infected. Proper steps must be taken to make sure data is coming from an authentic source and it is legitimate. Incoming data from End-Point devices such as smart phones is growing tremendously and filtering or validating data from these sources is a very big challenge [16].

4.6 Securing Transaction Logs and Data

Data in any organization many be stored at various levels (tiers) of the storage structure depending on the need and usage of the data. Increase in the transfer of data within the organization enforce for the need of auto-tiering for Big Data storage whereas auto-tiering does not maintain the log of where the data is stored and hence security is a big concern.

4.7 Securing Distributed Framework

Distributed framework enforces parallelism. This means that data is distributed across multiple nodes to achieve faster processing of large volumes of data. This increases the security concern of the framework and the data that exists there. Most companies use a distributed framework like MapReduce in which mappers read and compute and reducers combine the output from each mapper. If mappers are not secured, there is the chance of data being compromised [4].

4.8 Data Provenance

It is very important to know the original data that is coming to the platform so that we can better classify them. Data Origin should be consistently monitored but in reality due to the high volume it is becoming a big concern for data security. Provenance metadata is growing significantly as well and protecting metadata is very crucial for any organization [4].

5 BIG DATA SECURITY STAKEHOLDERS

In the digital era, the traditional way of securing the data, changing passwords frequently, firewall protection is just not enough to keep up with the growth of data produced by Internet of Things("IoT), Smart Devices, Bring Your Own Devices ("BYOD") and several customer friendly apps that is coming out everyday. "Even though

end user has the biggest responsibility with securing his own data, unfortunately, end users are not fully aware of the cyber security issues and they do not have the appropriate knowledge to discover the world wide web in complete safety" [17].

Big Data deployment is not possible to handle by any single business unit or with single tech team. It involves several business units, infrastructure, information technology, security,compliance, programmers, testers and product owners are all involved in big data deployment. They are all responsible for Big Data Security. Information Technology and Security team is responsible for drawing the policies and procedures. Compliance officers together with security team will protect compliance, such as automatically encrypting personally identifiable information before it is easily accessible. Administrators will automate these process to protect their environment. Even though every organizations have their policies and control laid in place to protect their biggest asset, phishing attacks can come in any form as a simple email. Frequent internal audit within the company can help us periodically check if all privacy, security and compliance are all in place.If not, proper measures can be taken right away to avoid any legal issues.

"The average annualized cost of cyber crime based upon a representative sample of 237 organizations in six countries by Ponemon Institute in their 2016 Cost of Cyber Crime Study and the Risk of Business Innovation sponsored by Hewlett Packard Enterprise is 9.5million U.S. dollars" [6]. In any organization, loss of information is the most expensive consequence of a cyber crime. The cyber attack may results in business disruptions, data or information loss, loss of revenue, damage to equipment and last but not the least it damages the brand. So it is big time to protect and secure the big data and the environment from all angles.

6 BEST PRACTICES FOR SECURING BIG DATA

There are three fundamental principles used in defining security goals: confidentiality, integrity, and availability. Confidentiality is the ability to keep sensitive data safe from third parties and unauthorized access. Integrity in this context means to avoid unauthorized modification of the data. Finally, availability means always being able to access the data and resources. These three concepts are known as the CIA triad, and is used as base principles when discussing and designing security practices [7].

To meet these goals, there are four main branches of security that apply to Big Data: Authentication, Encryption, Data Masking and Access Control [1].

6.1 Authentication

Because of its nature (large sizes of data, linking different sources, sharing access with third parties, etc), some of Big Data's features are highly susceptible to different privacy, security and welfare risks [9].

Privacy can be defined as the condition of confidentiality, protecting information from third parties. To support privacy, there have been different Authentication methods that both verify and validate entities who attempt to access the information. This ensures that only authorized entities are able to access the data or resources.

With Big Data, it is important to choose a proper authentication method, with the least computation complexity as possible, to allow dynamic security solutions within large Data Centers and also to avoid incrementing the traffic unnecessarily. Choosing an overbearing authentication method can cause both delay and storage issues. Because of this, it is important to tailor the security to the needs of the specific network [22].

6.2 Cryptography

There are multiple understandings of how data moves through stages, also known as Data Life cycles. Cryptography- define in terms of security. CITA.

From the perspective of cryptography, there are three phases in the Data Life Cycle: Data in Transit, Data in Storage, and Data in Use. Different cryptography techniques will be implemented depending on which stage of the life cycle the data is in [7].

There are different cryptographic tools that not only keep data secure at each point in its life-cycle, but also enable richer use of the data. The main tool is Encryption. Encryption takes pieces of data in plain text and use a cryptographic key to produce a version of the data that can only be read using the cryptographic key. Without the key, the information is illegible. There are two types of encryption: secret key encryption and public key encryption. Secret key encryption is when the same key is used for both encrypting and decrypting data. There are scenarios when one of the keys can be made public. For instance, if the locking key is kept private but the unlocking one is made public, this security can be used to prove authenticity [7].

There are different standards for encryption. The most well known and commonly used is Advanced Encryption Standard (AES). This standard sets guiding principles to ensure that data is encrypted in a manner that meets security needs and allows the recovery of original data [7].

6.3 Data Masking

By definition, Big Data works with large volumes of heterogeneous data sets using software to manage the data and to provide predictive analysis. Data masking works by replacing sensitive data with non-sensitive values, yet preserves the data integrity. For instance, replacing names with code names, or social security numbers with a key number. By doing this, different parties can access information without putting sensitive data at risk [2].

Five laws for data masking have been developed by Securosis Research. The first law is that data masking should not be reversible. This means that the data should not be unmasked easily using reverse engineering. The second law is that data that has been masked has to represent the original data set. For example, it has to belong in the same context. The third law states that data masking should maintain application and database integrity. This means that the process of data masking should not modify or affect the data in the databases in a negative way. The fourth law emphasizes that non-sensitive data can be masked, but it should not be masked if it can not be used to make sensitive data vulnerable. For instance, when masking information about a person, it is correct to mask the personfis name, email address and social security number, but other information like gender, or favorite colour, would be useless

to mask. Finally, data masking must be a repeatable process, using a standard to reproduce the steps taken to mask the data, allowing to troubleshoot possible problems in the process [2].

6.4 Access Control

As it was explained in the Challenges section, Access control, allows some entities to access the data or resources, while denying its use to other entities. Through security settings.

Some authors add that the inferences drawn from data should also be a cause for concern, because they can identify traits and patterns that could expose vulnerabilities. They propose that organizations who use the protected data should disclose their decisions criteria in order to apply access control in a broader spectrum. By doing so, it would be sufficient to diminish privacy concerns by de-identifying the data, or denying access to certain parts of the data that coul be used to make entities or data vulnerable. Some of these authors say that by doing this, it would not only reduce the privacy risk, it would also salvage large amounts of data for alternative use. This de-identification can also be achieved through data masking, pseudonymization, aggregation, among other methods [21].

6.5 Physical Security

It is always better to build and deploy Big Data platforms in their own data center. If deployed in a cloud, the organization must diligent to ensure that the cloud provider's data center is physically well secured. Access should be restricted to strangers and staff who have no official responsibilities in the designated areas or interacting with the data sources. Data centers should be properly monitored at all times and video surveillance and security logs are important tools to achieve this.

7 FUTURE OF BIG DATA SECURITY

To think about Future of Big Data Security, it is necessary to engage the conversation of what the trends are in Big Data and what technologies are expanding and changing the horizon. There are many new technologies and solutions that are shaping the future of the Big Data, but because of the length and focus of this document, there are three main areas that will be covered: Virtualization and Cloud Computing, IOT Security, External Password Vaults and Penetration Tests.

7.1 Virtualization and Cloud Computing

Virtualization is a way of deploying resources at multiple levels, such as hardware, network infrastructure, application and desktop centralized managing and using dynamically the physical resources. This makes the system flexible and less costly than traditional environments and giving management new tools to optimize the use of resources [14].

Since virtualization can be developed in so many levels, including cloud computing and by multiple service providers, it is natural that the system requirements of users and organizations move towards a variety of solutions that may include Infrastructure as a Service (IaaS) frameworks from public clouds such as the ones offered by Amazon, Microsoft, Google, Rackspace, HP, among others, or even Private clouds, maintained and many times even set up by internal IT departments [23].

These cloud computing technologies are being used to solve data-intensive problems on large-scale infrastructure. Thus, integrating big data technologies and cloud computing for data mining, knowledge discovery, and decision-making [10].

7.2 IOT Security

The Internet of Things (IoT) is the name given to the large network of physical devices that does not match the typical concept of computer networks, this includes all kinds of objects. The large and growing amount of devices and diverse uses given to them, makes IoT generates very important Big Data streams. Making it necessary to develop new systems and data mining techniques for this new paradigm [3].

In this IoT paradigm, each new opportunity opens doors to new threats as well. This makes it necessary to develop techniques to ensure trust, security and privacy. Different Authors write about the possible ways to face these challenges, and some, they consider three main axes to articulate the solutions: Effective security - used in very small embedded networks,context-aware privacy and user-centric privacy, and the third one is the systemic and cognitive approach for IoT security - where the interaction between people and the IoT can be envisioned as a set of nodes and tensions [18].

All this to say that in order to approach privacy and security in this new paradigm, many new theories and techniques have been developed since old security products and techniques may not suffice the needs of the different IoT users and communications.

7.3 External Password Vaults

Password vaults are applications that store multiple passwords and encrypt them storing them in a database [5].

There are small Password Vaults that can be stored locally on a system, or larger options that can be integrated into larger systems, providing additional security options, like generating real time temporary passwords for effective password rotation (I.E. Cyberark External Password Vault) [12].

These techniques are key to articulate authentication and a proper data access while using multiple services such as Cloud infrastructure and IoT.

7.4 Penetration Tests

After applying all the security techniques and strategies, and after putting in place all necessary security and privacy policies, the most important step is validating the strength of the security of the system. For some time, companies have started to perform tests that consist on simulating an attack from the perspective of an attacker, this method is known as Penetration test and it allows to actively evaluate and assess the security of a system [19].

The tester identifies the threats faced by an organization from hackers and suggest changes to improve the security and minimize the vulnerabilities and close the possible loop holes in the network [19].

8 CONCLUSIONS

Big Data as a constantly evolving and ever changing branch of information technologies resembles an ecosystem that since it covers gathering data from so many sources, processing it and generating

new information, there will be many entities and interests involved that will need to be protected.The features of Big Data such as Volume, Variety and Velocity bring new challenges to security and privacy protection. To protect the integrity and availability, security providers and local IT departments, will have to diversify their security and privacy strategies and policies, in order to keep pace with the growth and evolution of this new ecosystem.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Karim Abouelmehdi, Abderrahim Beni-Hssane, Hayat Khaloufi, and Mostafa Saadi. 2017. "Big data security and privacy in healthcare: A Review". *"Procedia Computer Science"* 113, Supplement C (2017), 73 – 80. https://doi.org/10.1016/j.procs.2017.08.292 The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.

[2] RA Archana, Ravindra S Hegadi, and TN Manjunath. 2017. A Big Data Security using Data Masking Methods. *Indonesian Journal of Electrical Engineering and Computer Science* 7, 2 (2017), 449–456.

[3] Albert Bifet. 2016. Mining Internet of Things (IoT) Big Data Streams.. In *SIMBig*. CEUR-WS.org, CUSCO, PERU, 15–16.

[4] Peter Buttler. 2017. *Big Data Needs Big Security. Herefis Why.* Technical Report. Dataconomy, Berlin, Germany. http://dataconomy.com/2017/07/10-challenges-big-data-security-privacy/

[5] Rahul Chatterjee, Joseph Bonneau, Ari Juels, and Thomas Ristenpart. 2015. Cracking-resistant password vaults using natural language encoders. In *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, CEUR-WS.org, CUSCO, PERU, 481–498.

[6] Ponemon Research Department. 2016. Ponemon Institute Research Report. (October 2016). https://www.ponemon.org/local/upload/file/2016%20HPE%20CCC%20GLOBAL%20REPORT%20FINAL%203.pdf

[7] Ariel Hamlin, Nabil Schear, Emily Shen, Mayank Varia, Sophia Yakoubov, and Arkady Yerukhimovich. 2016. *Cryptography for Big Data Security*. Taylor & Francis CRC Press, Boca Raton, Florida. 241–288 pages.

[8] A.J.G. Hey, S. Tansley, and K.M. Tolle. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, REDMOND, WASHINGTON. https://books.google.com.my/books?id=oGs_AQAAIAAJ

[9] Nir Kshetri. 2014. Big data s impact on privacy, security and consumer welfare. *Telecommunications Policy* 38, 11 (2014), 1134–1145.

[10] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya. 2016. The anatomy of big data computing. *Software: Practice and Experience* 46, 1 (2016), 79–105.

[11] Jose Moura and Carlos Serrao. 2016. Security and Privacy Issues of Big Data. *CoRR* abs/1601.06206 (2016), 1–2. https://doi.org/10.4018/978-1-4666-8505-5.ch002 arXiv:1601.06206

[12] Gregory S Nelson. 2015. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. In *SAS Global Forum Proceedings*. SAS Institute Inc., DALLAS, TEXAS, XXIII.

[13] Cheryl OfiNeill. 2017. *Big Data Needs Big Security. Herefis Why.* Technical Report. Imperva, Redwood Shores,California. https://www.imperva.com/blog/2017/02/big-data-needs-big-security-heres/

[14] K Padmini. 2015. Securing data management based on key technologies in cloud computing. *International Journal of Advance Research in Computer Science and Management Studies* 3, 2 (February 2015), 165–172.

[15] Amine RAHMANI, Abdelmalek AMINE, and Mohamed Reda HAMOU. 2015. A Mathematical MODEL OF ACCESS CONTROL IN BIG DATA USING CONFIDENCE INTERVAL AND DIGITAL SIGNATURE. *Computer Science & Information Technology* 5 (November 2015), 183–198.

[16] Sreeranga Rajan, Wilco van Ginkel, and Neel Sundaresan. 2012. *Top Ten Big Data Security and Privacy Challenges*. Technical Report. Cloud Security Alliance. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf

[17] Realdolmen. 2017. Cyber Security. (July 2017). http://www.realdolmen.com/en/blog/who-responsible-for-data-security-your-company

[18] Arbia Riahi, Enrico Natalizio, Yacine Challal, Nathalie Mitton, and Antonio Iera. 2014. A systemic and cognitive approach for IoT security. In *Computing, Networking and Communications (ICNC), 2014 International Conference on*. IEEE, ICNC, HONOLULU, USA., 183–188.

[19] Chaitra N Shivayogimath. 2014. AN OVERVIEW OF NETWORK PENETRATION TESTING. *International Journal of Research Engineering and Technology* 03, 07 (2014), 408–413.

[20] Christine Taylor. 2017. *Big Data Security*. Technical Report. Datamation, Foster City, California. https://www.datamation.com/big-data/big-data-security.html

[21] Omer Tene and Jules Polonetsky. 2012. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11 (2012), xxvii.

[22] Vijey Thayananthan and Aiiad Albeshri. 2015. Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. *Procedia Computer Science* 50 (2015), 149–156.

[23] Gregor von Laszewski, Fugang Wang, Hyungro Lee, Heng Chen, and Geoffrey C. Fox. 2014. Accessing Multiple Clouds with Cloudmesh. In *Proceedings of the 2014 ACM International Workshop on Software-defined Ecosystems (BigSystem '14)*. ACM, New York, NY, USA, 21–28. https://doi.org/10.1145/2609441.2609638

[24] Vitthal Yenkar and Mahip Bartere. 2014. Review on fiData Mining with Big Datafi. *International Journal of Computer Science and Mobile Computing* 3, 4 (2014), 97–102.

A WORK BREAKDOWN

Research for Section Big Data Needs Big Security, Big Data Security and Challenges: Uma Kugan.

Research for Section Best practices and Future: Andres Castro Benavides.

Editing: Andres Castro Benavides and Uma Kugan.

Big Data and the Issue of Privacy

Ashley Miller
Indiana University
admille@iu.edu

ABSTRACT

The collection, analysis, and dissemination of findings from big data can provide benefits to society across a variety of industries including health care, education, law enforcement, and commercial retailers, to name a few. However, with benefits, also comes caveats, risks, and even significant harm when this process of harnessing big data does not go as planned. One prominent discussion point around the use of big data is around privacy. In today's age of sharing information, can one really expect privacy given all the methods in which they can share information? What impact can data breaches have on a person's (or company's) personal information, security, and livelihood? While there are techniques and tools that have been implemented to help minimize the risk, is a data breach more a matter of *if* it occurs or more so *when* occurs? We will seek to address these questions as we explore how the increasing and ever changing space of online behavior, data mining, and big data analytics affects today's experience with maintaining privacy.

KEYWORDS

i523, hid329, big data, privacy, security, data mining

1 INTRODUCTION

As defined by Merriam-Webster, privacy is the "quality or state of being apart from company or observation" [11]. As we explore the world of internet behavior and big data, the question is: should we ever expect *privacy* as we are constantly being observed in this space? The Supreme Court has weighed in on this debate by stating that "one cannot have a reasonable expectation of privacy in information that is given to third parties or made accessible to the public" [12]. However, the government sector itself is one of the biggest data collectors across various agencies and hundreds, if not, thousands of databases [12]. With the vast amount of data collected by these agencies, ranging from political contributions to payroll, the definition of personal identifiable information (PII) can vary even across state lines as well as the laws and policies that govern them [1].

2 WHAT IS PERSONAL IDENTIFIABLE INFORMATION? (PII)

Despite the differences that may exist in what constitutes PII and what does not, there are main components these definitions have in common. PII is defined as an individual's name (first name and last name initial or the full last name) in conjunction with one of the following [1]:

- An individual's social security number
- A state identification or driver's license number
- Information related to an individual's protected health information, which could include health status or payment method
- A password to access an account such as a debit or credit card

While the Federal Trade Commission (FTC) provides guidelines and regulations in regards to storing personal information, they do not necessarily have an explicit definition for what constitutes PII [1]. Due to this, "reputation-implicating information" is not included [1]. However, this does not mean that companies are absent from blame when data breaches occur that inflict reputational harm. Fair information practices (FIPs) help to give guidance to organizations as they seek to balance the responsibility of collecting and maintaining personal information but also adhere to data security and privacy laws [7]. Despite this, the United States has no one law or set of laws that would require all organizations to follow FIPs as separate laws cover particular industries such as the Health Insurance Portability and Accountability Act (HIPAA) for health care organizations, Fair Credit Reporting Act (FCRA) for consumer credit agencies, as well as many others [7]. These separate set of provisions along with inconsistent, and even outdated, laws and practices from state to state, call into question further how one's privacy can be compromised.

3 HOW CAN PRIVACY BE COMPROMISED?

Cunan and Williams outline two key privacy issues in Figure 1 regarding information reuse and unauthorized access [7].

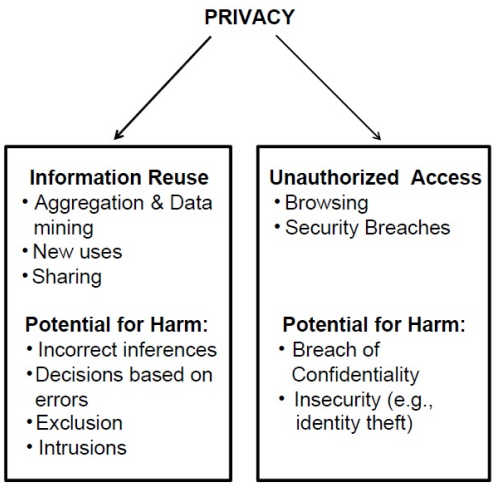


Figure 1: Privacy Issues

Data breaches would fall into the category of *unauthorized access.* As of information presented in 2009, it was then estimated

that nearly 250 million records had been exposed to some sort of data breach [4]. Given recent news of data breaches among major companies such as Yahoo!, Target, and Experian, recent figures now estimate that number to be 825 million records [13]. Data breaches can be described as a means to extort, expose or embarrass others [1]. Among a sample of 500 investigations, it was estimated that nearly nine of out ten data breaches could have been prevented [7].

Another way privacy can be compromised is when previously collected data is reused for a purpose in which it was not originally intended. A common data mining example cited is when Target used consumer purchasing information to predict pregnancy among consumers [16]. When coupons for baby and pregnancy related products were sent to a teenager's house, a local Target store received complaints from an irate father who only later found out that his teenage daughter was, in fact, pregnant [16]. While data mining and predictive analytics can provide benefits to consumers and companies, this one example brings up an interesting debate as to where the privacy line should be drawn when data is used for other purposes that a consumer may not be fully aware of or understand.

4 WHERE CAN PRIVACY BE COMPROMISED?

Like local retail stores such as Target or even online providers such as Amazon, consumers provide their data across a variety of industries, situations, platforms, and methods. Areas where PII could be compromised include but not are limited to:

4.1 Mobile

Smart phone applications alone can house a wealth of data among individual users [15]. In addition, sensors such as microphones, cameras, and global positioning systems (GPS) can offer more information as to what a user sounds like, looks like, and even their travel patterns [15]. It's hard to say whether consumers are fully aware of these possible data collection methods and what measures, if any, they take to protect their privacy on a mobile device such as turning off location finding features or password protecting accounts. However, a study conducted by Dimensional Research shows that one out of every five security professionals surveyed have had their company experience a data breach from their mobile device while nearly one out of four state they are unsure if one has occurred [6].

4.2 Health Care

A patient's family history, prescriptions taken, diseases, and disorders are stored by physicians' offices and can live in electronic medical records (EMR) across the globe. Health insurance companies are also another player in collecting and maintaining big data on their consumers. While harnessing this information can provide substantial benefit to developing outcomes and even identifying trends, it can also be at risk of being compromised if not properly secured and maintained [15]. One such example includes nearly 80 million records that were compromised from Anthem in 2015 [2]. Considering these records can contain home addresses, social security numbers, and even names of partners and children, the damage one can inflict to an individual can be extensive and overall, the impact of these data breaches are said to have cost the health care industry nearly 5.6 billion dollars per year [2].

4.3 Education

Even in the realm of education, individual information around academic performance, class behavior, as well as intellectual disabilities can be at risk of being compromised [3]. Higher education systems in particular can at times lag behind other industries in their attempt to maintain privacy and security [8]. Two large data breaches occurred among Indiana universities. One occurred at Butler University in 2014 that affected nearly 200,000 people and sensitive information leaked ranged from social security numbers to bank accounts [10]. Indiana University also experienced a similar size of data breach that cost the school approximately 130,000 dollars [10].

4.4 Online Accounts

Across social media, search engines, and other websites, massive amounts of data are stored on individuals [15]. Facebook by itself houses information on more than 900 million users which can includes photos, friend lists, likes, shares, and as well as personal information provided such as name and location [15]. Browsers and operating systems collect online and mobile information about users and also can track information across search history [15]. Online commerce takes up a large percentage of the retail market [15]. With this, there is an increased number of transactions taking place online which opens the doors to buyer history as well as possible fraudulent activity among hackers and scammers who wish to steal payment account information for their own benefit [15].

5 HOW CAN PRIVACY BE BETTER PROTECTED?

Given the on-going need to understand consumers through data, there are ways to further protect privacy among individuals, including but not limited to:

5.1 Deidentification of data

While not new to data mining and analytics, the ability to anonymize, encrypt, or code the data in a way which removes PII can help to protect one's privacy [14]. Even with this process, there is always the risk that a user may be *re-identified* if multiple data points are then grouped back together that give specifics about an individual [9]

5.2 Privacy Preserving Data Mining (PPDM)

This method highlights a two-prong approach where sensitive data, such as an identification number or phone number, is removed and not used for data mining purposes [16]. Further, any results produced that could violate privacy are excluded [16]. This approach in Figure 2 details user roles [16]. Within each user role, privacy concerns could exist among those involved in big data mining:

- **Data Provider**: This would be considered a consumer who may be sharing their information online with a company

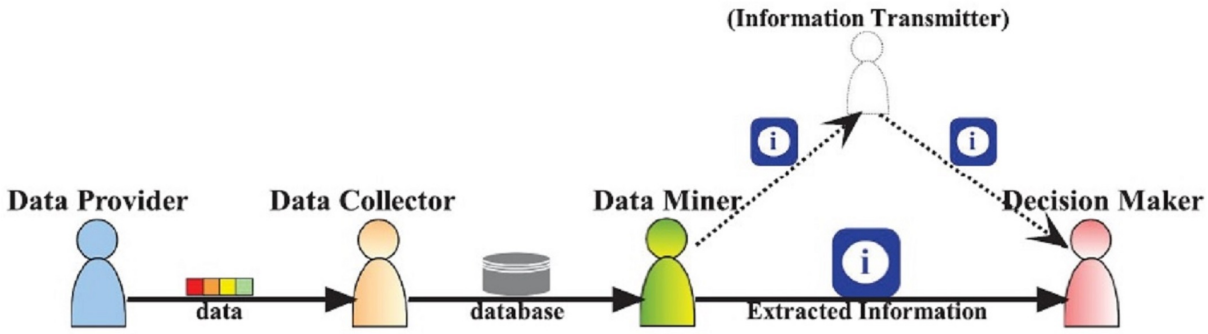


Figure 2: Data Mining Roles

during a transaction. Here, concerns around control can affect behavior [16]

- **Data Collector**: Those who collect the users' information want to ensure that the value of the data provided is worthwhile without compromising privacy or security [16].
- **Data Miner**: During the process of yielding insights, this role's main concern is to keep sensitive information from unintended parties [16].
- **Decision Maker**: Application, ownership, representation, and storage are privacy issues among decision makers as they seek to understand insights from the sources and people provided [16].

Given these concerns, there are also instances where maintaining the original data may be needed. Another approach is reversible privacy preserving data mining (RPPDM) [5]. This technique still preserves data privacy but also allows for a restore procedure to take place in case there are issues that occured during the data distortion process [5].

5.3 Individual Access, Control, and Responsibility

Giving users an option to *opt-in* or *opt-out* of how their data is collected and used may encourage consumers to play a more active role in their online privacy. Informed consent can provide benefits and drawbacks to both companies and consumers [14]. Adherence is often higher in instances where users *opt-out* of certain features or functions [15]. However, increased emphasis on obtaining consent can also lead to fewer innovations and benefits to society at large if consent has to be obtained for big data collection [15]. Giving someone the ability to access their own information can alleviate privacy concerns but consumers are often unaware of this possibility [14].

Another option to consider is similar to that adopted by the European Union (EU), as one's *right to be forgotten* where users can have their information removed from websites [9]. Tools and other website information may also be available to give the users more control of their data. One tool, personal.com, allows users to gain access to personal information of theirs and control use [15]. Unless individuals have a vested interest in their own privacy management, then little improvement can be expected [14].

5.4 Increased Onus on Companies and Organizations

While laws and regulations can only do so much, increased pressure by government and consumers may in fact help to entice these companies and organizations to improve their own policies and procedures in a more proactive manner [7]. Ultimately, these changes have to come from a leadership level to be effective across the organization. While one could argue doing so is part of doing *good business*, these additional measures can increase costs for organizations [7]. Areas to consider could include cross-functional privacy sub-committees, privacy impact assessments (PIA), and cyber liability insurance [7].

5.5 SenseMaking (Privacy by Design System)

With disparate data sources being prevalent in any organization, sensemaking allows for observations to live together [4]. This analytical big data platform was designed with privacy in mind from the ground up with key features that cannot be disabled [4]. Features outlined by the system include elements related to full attribution, data tethering, anonymized data analytics, audit logs that are resistant to tampering, methods that favor false negatives and correct for false positives, in addition to transfer accounting of information [4].

6 CONCLUSION

As technological capabilities, level of knowledge, personal accountability, and experience evolve, so may the level of security and process that maintains and protects one's personal identifiable information (PII). However, the possibility of data being compromised may always exist. While a consumer can take measures on their own to better protect their privacy, companies and organizations have to also be a partner in the data privacy preserving methods if

they wish to use big data mining and analytics methods to their advantage. This effort also helps to keep costs down as data breaches can be very costly not only to an organization's infrastructure and resources but also to one's reputation in their effort to conduct good business with their constituents. There is no one method in particular to show how to achieve the appropriate balance of big data mining and privacy but continued effort in learning from big data breaches will help to ensure that they occur less often and with less harm to others when they do occur.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the teaching assistants for their support and guidance in writing this paper.

REFERENCES

[1] Yasmine Agelidis. 2016. Protecting the Good, the Bad, and the Ugly: "Exposure" Data Breaches and Suggestions for Coping with Them. *Berkeley Technology Law Journal* 31 (2016), 1058–1078.

[2] Nsikan Akpan. 2016. Has Health Care Hacking Become an Epidemic? (2016). https://www.pbs.org/newshour/science/has-health-care-hacking-become-an-epidemic

[3] Justin Bathon. 2013. How Little Data Breaches Cause Big Problems. *T H E Journal (Technological Horizons In Education)* 42, 8 (2013), 26–29.

[4] Ann Cavoukian and Jeff Jonas. 2012. Privacy by Design in the Age of Big Data. (2012).

[5] Tung-Shou Chen, Wei-Bin Lee, Jeanne Chen, Yuan-Hung Kao, and Pei-Wen Hou. 2013. Reversible Privacy Preserving Data Mining: A Combination of Difference Expansion and Privacy Preserving. *Journal of Supercomputing* 66, 2 (2013), 907–917.

[6] Stacy Collett. 2017. Five new threats to your mobile security. (2017). https://www.csoonline.com/article/2157785/data-protection/five-new-threats-to-your-mobile-security.html

[7] Mary Culnan and Cynthia Clark Williams. 2009. How Ethics can Enhance Organizational Privacy: Lessons from the ChoicePoint and TJX Data Breaches. *MIS Quarterly* 33, 4 (2009), 673–687.

[8] Ben Daniel. 2015. Big Data and Analytics in Higher Education: Opportunities and Challenges. *British Educational Research Association* 46, 5 (2015), 904–920.

[9] John G. Francis and Leslie P. Francis. 2014. Privacy, Confidentiality, and Justice. *Journal of Social Philosophy* 45 (2014), 408–431. Issue 3.

[10] Kyle McCarthy. 2015. 5 Colleges With Data Breaches Larger Than Sonyfis in 2014. (2015). https://www.huffingtonpost.com/kyle-mccarthy/five-colleges-with-data-b_b_6474800.html

[11] Merriam-Webster. 2017. Privacy. (2017). https://www.merriam-webster.com/

[12] Harvard Law Review. 2014. Data Mining, Dog Sniffs, and the Fourth Amendment. (2014), 691-712 pages. Issue 2.

[13] Adam Shell. 2017. Equifax Data Breach: Number of Victims May Never Be Known. (2017). https://www.usatoday.com/story/money/2017/09/17/equifax-data-breach-number-victims-may-never-known/670618001/

[14] Omer Tene and Jules Polonetsky. 2012. Privacy in the Age of Big Data: A Time for Big Decisions. *Stanford Law Review Online* 64, 63 (2012), 63–69.

[15] Omer Tene and Jules Polonetsky. 2013. Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property* 11 (2013), 240–272. Issue 5.

[16] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren. 2014. Information Security in Big Data: Privacy and Data Mining. *IEEE Access* 2 (2014), 1149–1176.

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 a 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 aperature 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 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

[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 Throwable 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.c4isrnet.com/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

Security aspect of NOSQL databases in Big Data Application

Budhaditya Roy
Indiana University
School of Information and Computing
Bloomington, IN 47040
royb@indiana.edu

ABSTRACT

NOSQL databases play an integral part in analyzing and organizing vast quantities of data. The recent advancement of big data applications involve datasets which are fast changing, massive and diverse in nature. As innovation of big data progresses, next big thing comes to the world of development is where to store this voluminous data which is core of any big data application. With the evolution of NOSQL databases data storage problem was resolved but a new concern has risen sharply, the fiSecurityfi. In todayfis world data governance, in form of data security plays a most imperative role in success of every organization. NoSQL databases are designed to deliver real time performance in keeping large volume of data stored but while developing these databases security was not a primary subject rather performance, velocity, scalability were top of the list. This paper talks about the security aspect of NOSQL databases in big data application and how organizations can implement NOSQL databases considering the security and whether having these databases a wise decision on picking up right databases in big data application.

KEYWORDS

NoSQL databases, Security, HID348, Big Data Applications in NOSQL, MongoDB, Apache Cassandra, Confidentiality, Integrity, Availability

1 INTRODUCTION OF NOSQL IN BIG DATA

In last 20 years we have seen the data boom where the volume, velocity and variety of data has increased almost nine times and within last five years it has become even more. Big Data refers to collection of large volume of data characterized as multi fiVfi [6]. Big Data and Data intensive technologies are going through a technological advancements with relation to all aspect of human activity [3]. NoSQL, fiNot Only SQLfi is a non-relational databases which is certainly popular databases which are scalable, help in large big data application deployment, easy to implement and highly usable in storing unstructured and semi structured data. NOSQL databases are also very cost efficient and most of the time these are open source. When scale this solution, the cost factor not always gauge with security. This databases are different. NoSQL is designed to be accessed by trusted clients. NoSQL databases are flexible databases used in big data applications and real time web apps. These kind of databases do not have a predefined schema and a flexibility in data model are the feature which can be a great benefit for the companies to implement this product[6]. NOSQL also has not predefined data structures along with ability to handle huge amount of unstructured data. NOSQL databases also have remarkable benefits in scaling, it uses scale out/horizontal scaling methods whereas traditional relationship databases use vertical scaling or scaling up. In todayfis world 90 percent of use cases are not required a relational database, RDDMS is often implemented because of support from IT team and which can be easily productionize. Since successful applications are gaining more users and more data in every second scaling become an essential part of big data industry. Along with aforesaid benefits NOSQL databases carry a significant security risks where compromising data is possible occurrence [2]. There have been studies on whether NoSQL databases can stand alone in the organization and truly used against relational databases keeping in mind the importance of data.

2 SECURITY ISSUES IN NOSQL DATABASES

There are many security issues in big data. The most important security issue is data protection and access control[2]. NoSQL databases are great for big data but security is repeatedly lacking in NOSQL applications. There is a high need of having access control on the semantic layer of the data as well as in place superior attribute relationship methodology. As NOSQL databases are designed to provide real time performance while managing large volume of data there is a risk of security implication moving from relational databases to NOSQL database. Currently NOSQL databases are in evolutionary stage of their lifecycle and as they progress daily, security protocols for NOSQL databases are not well defined yet. Days are probably not far when we would see a data breach from NOSQL injections[2]. Question is whether NOSQL is really that vulnerable to security breach. There are two aspect of answering this question. Firstly, NOSQL databases are designed to hold gigabytes of data which is a golden ball to attackers, secondly, since NOSQL databases are primarily developed for performance impressive and as high scalable product security was not at all primary. Many of the NOSQL products recommend users to have the TCP/IP in trusted environment but in an internetfi?!enabled world relaying on network would be too dangerous to have. Security measures are mainly divided into four main categories such as Injection, Authentication, Authorization and Confidentiality.

2.1 Injection Attack-

Every databases are built on certain languages. To attack any databases it is required to know which language is used. SQL injection attacks are increasing daily and NOSQL databases are mostly vulnerable to these attacks. Since NOSQL databases do not use SQL, instead use JavaScript Object Notation (JSON) query language and HTTP API that makes traditional injection obsolete. In recent times Schema injections are used most frequently to use as an extra protection layer. There are two type of schema injection

overrides can be implemented. First is override schema in JSON object itself and another is external override schema [3].

2.2 Authentication-

Authentications are most important aspect of any databases irrespective of the type, especially in NOSQL database, there is a need of strong authentication channel, strong password protection methods and password bruteforcing opportunism. In NOSQL authentication was not enables at all when starting newly established application. Two types of authentication methods which strongly need to be enables in NOSQL databases are, HTTP/RESTful authentication and Non-HTTP/RESTful authentication. Databases such as Apche CouchDB uses one of three types of authentication HTTP Basic, DIGEST or Coockie based authentication [3]. There is a lack of usability of SSL and encryption in these kind of databases where HTTP authentications are used. A reverse proxy server or loan bouncer is advisable for these databases.

2.3 Authorization-

In relational database Data control language or DCL plays a secure role in table level security measures. This is native access control built in to the language. In NOSQL there is no such DCL concept available right now except for database level access control it is more architecture dependent [5]. Most NOSQL databases have some common authentication feature as Admin role which is related to DML access role in relational database. In general authorization is not required unless it is enabled for almost all NOSQL databases which can pose to data security risk.

2.4 Confidentiality-

There is a lack of confidentiality exists in NOSQL database architecture. There is some SSL support but beyond that there is no extra SSL layer to support confidentiality[3]. There are third party software companies available which can provide support by adding a proxy servers to encrypt the data and users information, added to an extra cost.

3 OTHER SECURITY ISSUES OF NOSQL DATABASE-

Most of the NOSQL databases currently used have a very thin security layer along with string clustering mechanism which create more challenges in security implementation. There is a risk of security breaches from insider attack due to poor logging and log analysis method. There is also auditing risk in NOSQL security along with risk of database attacks when data at rest, in motion and in use stage.

4 SECURING NOSQL ENVIRONMENT-

There are couple of ways we can secure NOSQL deployments. There needs to be a trusted operating environment in place and there are necessities to understand on how the data are input and output from the system. There has to be access control on SSL encryption. Since these databases cannot operate on their own and often times these services are run on public IP addresses. There is a prerequisite of adding additional VPN into the architecture and increase TCO deployment times by adding an additional security layer. Besides, NOSQL environment need to be tight on validation. Since the NOSQL injection attack surface is diverse schema, JavaScript and query injection attacks affect NOSQL architecture differently. In order to prevent these attacks, it is necessary to understand how these attacks can affect the application as a whole along with NOSQL environment. There has to be a continuous validation to non-traditional injection attacks. Last but not the least there is a constant communication to the NOSQL vendor is highly opted in. Since NOSQL vendors have frequent releases, adding features to the NOSQL system is necessary to keep the environment secure.

5 NOSQL DATABASES AND SECURITY FEATURES-

Based on data storage model, NOSQL databases are categorized in following four sub categories. Such as, Key- Value databases, Colum databases, Document Databases, Graph Databases [3]. Main security feature of some of the NOSQL databases are discussed.

5.1 MongoDB

MongoDB is a document database with high performance, large scale high availability, and robust system. It is designed to run on top of data driven applications high level programming models, computing resources and process of automation. Trusted environment is the default option and is recommended. It is often better to run the database in a trusted environment with no in-database security authentication [3] . All data in MongoDB are stored as a plain text and overall environment lacks of data encryption [3]. Since MongoDB does not provide automatic encryption, attackers can easily access database files to extract information. Binary wire level protocols are not well connected to the client causing a lack of authentication support. MongoDB architecture is built on JavaScript language which is more prone to attack due to being an interpreted language. Even further MongoDB does not support data validation and data auditing and since authentication information is hash encrypted in MD5 algorithm there is a potential risk due to MD5 less tight security measures [1].

5.2 Apache Cassandra

Apache Cassandra Is a column based database which is distributed storage system. Files in Cassandra are kept unencrypted and there is no mechanism of automatic data encryption. There is a potential security compromise when database and client communicate due to lack of encryption as well. Cassandra has its own Cassandra query language or CQL which is disposed to to external security injection attacks [1]. Though Cassandra provides an encrypted intercluster network communication where enabling this feature is required from an external client.

5.3 Redis

Redis is an open source key-value store which is designed to be accessed by trusted clients inside trusted environment. [1] There is a key value match in these type of databases and data is stored in the key value pair. Which means it is not a inordinate idea to expose Redis instances directly to the internet where untrusted clients can directly access TCP/IP port and external intrusion is

very much conceivable. Network security is highly desirable for this environment where access to Rediss port should be denied to any kind of external access point preventing with a firewall. Redis is hard to protect from being accessed by external networks and many instances are exposed to public IP addresses [3].

5.4 Apache HBase

Apache HBase- Apache HBase is an open source column based database model. HBase is scaled to handle millions of data sets and billons of column and rows in form of unstructured and semi structured format by using wide variety of different structures and schemas [7]. Data security of HBase depends on SSH for internode communication [7]. It has security and authentication layer added to an extract security protocol.

5.5 Neo4J

Neo4J is an open source graph databases which uses SSL protocol to communicate between database and client. There is no data encryption between database server and client which allows potential security vulnerability. It provides node level data security based on ACLfis with groups, users and variety odd access levels. Though user authentication is prevailing in Neo4J, there is a lack of security in overall database level [3].

5.6 Apache CouchDB

Apache CouchDB is a document databases which allows any request to be made by anyone, any rogue client could enter along and delete a database [4]. Default installation in the interest is compromised. CouchDB supports authentication on coockie and password but there is no encryption in database level as well as client server commutation level. Authentication in this environment is only at database level and access control accepts only single user role authentication [4].

6 CONCLUSION

With the growth of big data, organizations more into NOSQL databases where security is a growing concern. Though we found there are severe security issues in most of the NOSQL databases which are used today in big data environment. Lack of security measures put extra sensitivity to the overall big data applications being NOSQL databases are heart of any big data project. Though not reached at pick, constant evaluation and research are in process to make NOSQL databases more secure in near future.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and I523.

REFERENCES

[1] ANA-MARIA BACALU ALEX POPESCU. 2012. *NoSQL Databases and Security: Cassandra and MongoDB Security Reviewed.* Web, Chapter 1, 2. http://nosql.mypopescu.com/post/15308448223/nosql-databases-and-security-cassandra-and

[2] Adarsh.G Dadapeer, N.M.Indravasan. 2016. *A Survey onSecurity of NoSQL Databases* (1st. ed.). 4th, Vol. 4. International Journal of Innovative Research in Computerand Communication Engineering, Ballari, Karnataka, India, Chapter 2. https://doi.org/10.1007/3-540-09237-4

[3] Mohammad Ali Nematbakhsh Ebrahim Sahafizadeh1. 2015. . 4, Vol. 4. Advances in Computer Science: an International Journal, Chapter 1, 5. https://pdfs.semanticscholar.org/e860/c0b19aa029008a6793acc8c293b68942234b.pdf

[4] Yunhao Liu Min Chen, Shiwen Mao. 2014. *Big Data: A Survey.* Springer Science+Business Media, NY, Chapter 1, 1. https://doi.org/10.1007/3-110-0825-3

[5] Abe Wiersma M.W. Grim. 2017. *Security and Performance Analysis of Encrypted NoSQL Databases.* 1, Vol. 1. University of Amsterdam.

[6] Purdue University Researchers (Ed.). 2015. *Challengee and Opportunities with Big Data.* 1st, Vol. 1. Purdue University Press, Purdue University, Indiana. https://doi.org/NA

[7] Web. 2016. *Apache HBase.* Chapter 1, 1. https://doi.org/10.7/3-105.876

Big Data Applications in Team Sports Predictive Analytics

Josh Lipe-Melton
Indiana University
4400 E Sheffield Dr
Bloomington, Indiana 47408
jlipemel@umail.iu.edu

ABSTRACT

We discuss Big Data Applications in predictive sports models. The models examined include genetic tuning, neural networks, and "per possession" models, which all take various forms of big data about a sports match and use it to make some sort of prediction about future matches.

KEYWORDS

sports, analytics, predictive, neural network, HID105, I523

1 INTRODUCTION

The sports industry produces a lot of money, and because of this, analysis of sporting events has become increasingly explored field. As technology improves, more and more data is generated regarding sporting events. Therefore, there have been numerous attempts to create functions that predict the outcomes of sporting events. Many efforts attempt to create neural networks to model these outcomes. Others use genetic pruning algorithms. Others break sporting events down into possessions and create a "per possession" model to predict the points scored by each competing party. Still more use a combination of all of these. By using various statistics, the results of a significant number of games can be predicted.

2 EXPECTED GOALS MODEL

Arguably the most common method of predicting the results of soccer games is to create a prediction of the number of goals scored by each team . The result of subtracting these two numbers gives not only a prediction of which team will win, but an inherent level of confidence proportional to the difference of each predicted number of goals [4]. This model creates an "expected goals value" by predicting the number of shots and assigning each of these shots a value. These values are based on attributes such as angle from the goal, distance to the goal, body part used to take the shot, what type of approach was used to obtain the shot (dribble, short pass, long pass, etc.), and even the relevant FIFA video game ratings of the player taking the shot. Each value represents the predicted likelihood of scoring, with 0 being an impossible shot and 1 being a sure goal. By summing these values and incorporating the FIFA rating of the opposing goalkeeper, an expected goals value for a team is obtained. This model is able to predict the number of goals scored by each team about 20 percent of the time. The correct result of the match was found about 56 percent of the time[4].

[Figure 1 about here.]

2.1 Bivariate Expected Goals Model

A flaw with the previous example of an expected goals model is that it accounted only for the attack team's ability in its goal predictions. Apart from the ability of the goalkeeper, there is no accounting for the defensive ability of an opponent in prediction of expected goals. In a different model, defensive ability and attacking ability are both incorporated. The authors of this method created their model based on the idea that the goals scored two competing soccer teams are negatively correlated with one another. By using a bivariate Poisson model for soccer data, the authors created predictions for the number of goals scored by each team in a given match, and therefore the results of each game[1]. The covariates used in the bivariate Poisson regression model include: GDP per capita, population, home advantage, bookmaker's odds, market value, number of Champion's League players, number of club teammates, and the age of the coach. By running 1,000,000 simulations on the European Championships in 2016, predictions for each match were created, along with odds for each team to reach each round of the tournament. The odds of the model outperformed bookmakers' odds 42.22 percent to 39.23 percent in predictive accuracy[1]. The authors used their model in placing equal bets on every bet in the tournament with the service that provided the most favorable odds to the outcome predicted by their model. In doing so, they obtained a return of 30.28 percent after the tournament.The authors concluded that the scores of two soccer teams are indeed negatively correlated and that this is a sound notion to base a predictive model on [1].

3 NEURAL NETWORK METHOD USING PAST MATCH RESULTS

Another method of prediction solely uses past results to predict future results. In this method, a predictive model is based on the intuitive proposition that if team 1 has won their previous few games, team 2 has lost their previous few games, and team 1 has beaten team 2 the last two times they have played, team 1 will beat team 2[2]. The model proposed in this article assigns a value in the range [-5, 5] to the last five games played by each team as well as the last two games played between the two teams. The higher the number, the bigger the win. The lower the number, the bigger the loss. The predicted result of a game is a function of these numbers. Through a combination of a fuzzy logic table and a neural network algorithm, a result is predicted. First, the authors created a table with every possible value of x1-x12. Each of these combinations was then associated with a predicted result and a weight in the interval [0, 1] that indicated the confidence in the predicted result. These initial confidence intervals were then tuned The predicted result is drawn from the range [Big loss (BL), Small loss (SL), Draw (D), Small win (SW), Big win (BW)][2]. Using a sample size of 1056

matches, the source" assigned weights to the nodes in the neural network as in Figure 2. The trained model was applied to 350 results from other seasons and was correct when predicting a big loss 91.4 percent of the time, a small loss 83.3 percent of the time, a draw 87 percent of the time, a small win 84 percent of the time, and a big win 94.6 percent of the time[2]. This model greatly outperforms the previous model examined. However, the authors do cite flaws that come from not considering factors such as injured or suspended players, refereeing, or weather conditions [2].

Furthermore, this method's already impressive predictive accuracy could also be improved by taking into account strength of schedule, as a team that has narrowly won its last five games against very weak opponents would be favored against a team that has narrowly lost against very strong opponents. The machine learning techniques implemented in this study could have been improved by incorporating opponents' results into the model, giving more weight to wins against good teams.

[Figure 2 about here.]

4 NCAA ANALYSIS

In college basketball, the committee that decides who gets into the NCAA tournament makes use of a ranking system called Ratings Percentage Index, or RPI. RPI weights .25 of a team's ranking on their win percentage, .5 on their opponents' win percentage, and .25 on their opponents' opponents' win percentage. [5] This system is designed to encourage teams to schedule difficult opponents, as a large portion of the rankings is based on strength of schedule. This formula has significant influence on where teams are ranked. Unfortunately, "the RPI lacks theoretical justification from a statistical standpoint." [5] In general, it is believed that the model places too much emphasis on strength of schedule and not enough on performance. Attempts to utilize an improved version of this model have made an impact on seeding in college soccer and baseball as well. In these sports, wins are weighted to give more ranking points to an away win than a home win.[5] These types of alterations, however, do not address the fact that 75 percent of this ranking comes from a team's strength of schedule. This type of bias favors teams that are in strong conferences, even if they have poor records in their conference.

4.1 Per Possession Analysis

A proposed alternative to RPI is to use a "per possession model," or a model that predicts outcomes using statistics that are used in the context of efficiency with possessions. For example, offensive efficiency is found by dividing points scored by possessions and defensive efficiency is found by dividing points allowed by possessions citation[3]. These statistics are then used to calculate an offensive efficiency adjusted by the perceived strength of the opponent. Adjusted offensive efficiency, for example, is calculated by multiplying offensive efficiency by the average national offensive efficiency then dividing this number by the adjusted defensive efficiency of an opponent citation. By combining these adjusted efficiencies with other factors such as home court advantage, the authors made several models which created an estimation for "win probability," which can in turn be used to predict individual match outcomes or create a ranking system. By using win probability, the study we examine created models based on decision trees, rule learners, artificial neural networks, naive Bayes, and ensemble learners citation. The neural network and naive Bayes models were the most effective models, both predicting outcomes with about 72 percent accuracy. A surprising observation from the authors is that simpler models tend to work better than more complicated ones citation. Similarly, attempting to incorporate more features into the models tended to decrease predictive accuracy citation. The authors believe that there is a "glass ceiling" when it comes to accuracy predicting sporting events of around 74 percent citation. Each of these models is unable to predict any individual season at a rate greater than 74 percent.[3]

5 CONCLUSION

Basic statistics are being used in all sorts of analysis of sporting events. Advances in machine learning make it possible to create highly effective models to predict the outcomes of sporting events based on past performances. One model examined used numerous statistics as inputs to predict the number of goals scored by a certain team, and barely predicted results better than random selection. Another used advanced statistics based "pace of play" and the efficiency of each team involved. Using these statistics, the authors concluded that simplest models worked best, and predicted the results of games correctly about 72 percent of the time. Finally, a model that used simple inputs, the previous few results of either teams, into a neural network provided stunning accuracy in predicting the results of games. This reinforces the notion that simple inputs, especially those involving neural networks, provide the greatest accuracy in predicting the outcome of sporting events.

6 ACKNOWLEDGEMENTS

I would like to thank Dr. Gregor von Laszewski and Juliette Zerick for providing valuable feedback on my paper

REFERENCES

[1] A. Mayr A. Groll, T. Kneib and G. Schauberger. 2016. On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016. (2016). Retrieved Nov 1, 2017 from http://eprints.kingston.ac.uk/39162/1/MathSport2017Proceedings.pdf#page=166

[2] M. Posner A. P. Rotshtein and A. B. Rakityanskaya. 2005. FOOTBALL PREDICTIONS BASED ON A FUZZY MODEL WITH GENETIC AND NEURAL TUNING. (2005). Retrieved Oct 30, 2017 from https://link-springer-com.proxyiub.uits.iu.edu/content/pdf/10.1007%2Fs10559-005-0098-4.pdf

[3] Jesse Davis Albrecht Zimmermann. 2013. Machine Learning and Data Mining for Sports Analytics. (2013). Retrieved Oct 30, 2017 from https://lirias.kuleuven.be/bitstream/123456789/424505/1/CW650.pdf

[4] H.P.H. Eggels. 2016. Expected Goals in Soccer: Explaining Match Results using Predictive Analytics. (2016). Retrieved Oct 30, 2017 from https://pure.tue.nl/ws/files/46945853/855660-1.pdf

[5] Wikipedia. 2017. Rating Percentage Index. (2017). Retrieved Oct 30, 2017 from https://en.wikipedia.org/wiki/Rating_percentage_index

List of Figures

ORTEC	FIFA	Inmotio
Context	Player quality	Number of attackers in line
Part of body	Goal keeper quality	Number of defenders in line
Dist to goal		Distance nearest defender in line
Angle to goal		Distance goal keeper
Originates from		
Current score		
High		

[5]

Figure 1: Features of Expected Goals Model

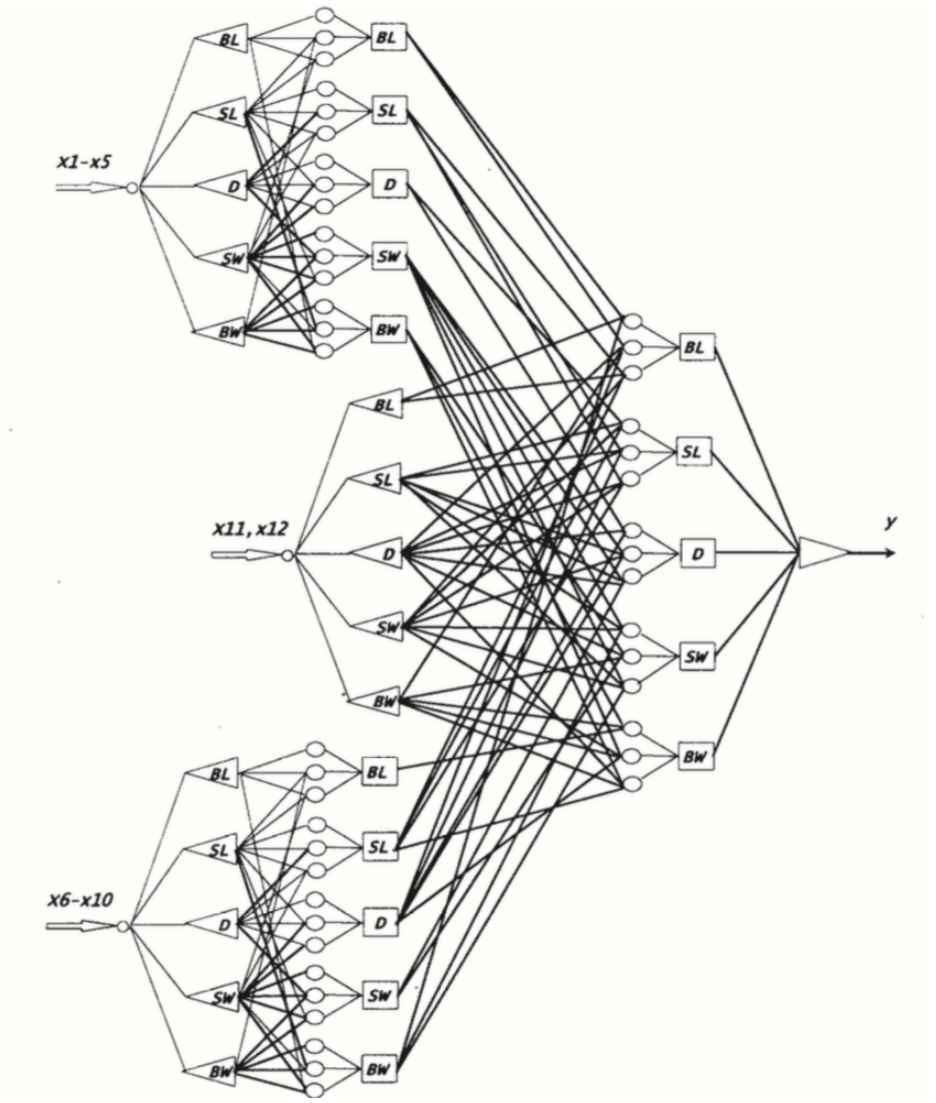


Figure 2: Depiction of the neural network algorithm. x1-x5 and x6-x10 represent the results from each team's last five games, while x11 and x12 represent the results from the previous two games between the teams

Nosql Database In Support Of Big Data Applications And Analytics

Juan Ni
Bloomington, Indiana 47401
nijuan@iu.edu

ABSTRACT

SQL database is the most usual database we use, the constitution of sql style database is easy for us to do data storage and search. According to the Artur's testing [1], he using many differences sql and nosql database to testing the insert performance for large amount of data(one million rows of randomized string) in the same time, and nosql databases always has best performance when insert large data. So when we need to deal with big data, SQL database is no longer handy for big data application. Even nosql database doesn't fit into every situation, but the mass data produce ability allow nosql database support most of big data applications and analytic.

KEYWORDS

523, HID 107, paper2, big data, database, SQL, NoSQL

1 INTRODUCTION

SQL and NosQL database have same goal which is to storage data, but they have different strategy about storage data by difference storage mode. As we know, SQL database using tables to storage data, the primary key and foreign key allow user making connection between several different tables. Table type storage mode is well organize, but it is less flexibly for some situation. For example we have a table which content two columns "Item"and"price", and one item's price still negotiate with the retail, so we want to put "Processing" at the price area; but the price area already set up only allow user import number data, so user can't make any flexible change to import data into table for meet the business need. Unlike SQL database only have one type of storage model, nosql database have four differences storage model based on the data categories, they are "Column, Document, Key value, and Graph"[10]. For example, nosql using key value which similar to JOSN to storage data, user can import any type of data inside the file, and the same type of file will storage as a collection. According to the nosql storage definition, the main differences of storage model between sql and nosql database is "Compared to relational databases, for example, collections could be considered analogous to tables and documents analogous to records. But they are different: every record in a table has the same sequence of fields, while documents in a collection may have fields that are completely different." [10]. Nosql supporting variable types of data storage, this advantage make nosql more suitable for big data application. The following section will discuss why nosql database have better performance for big data applications than traditional sql database, the advantage and drawback of nosql database.

2 MYSQL DATABASE VS NOSQL DATABASE

The above examples show that nosql database is much better than SQL database, but nosql database still has shortage at some domain compare with sql database. For example, the formatting of the table in SQL database has really strict data schema constraint, user only can import data by following the rule that already set up in the table, and this make sql database steady and rarely have error at data storage. The flexibility of nosql data storage might cause error happen according to jepsen's idea, "In this post, we'll see that Mongo's consistency model is broken by design: not only can "strictly consistent" reads see stale versions of documents, but they can also return garbage data from writes that never should have occurred. The former is (as far as I know) a new result which runs contrary to all of Mongo's consistency documentation. The latter has been a documented issue in Mongo for some time. We'll also touch on a result from the previous Jepsen post: almost all write concern levels allow data loss." [4]. Even the reliability of nosql is not good enough as sql, but nosql still better for big data application because of its cost-performance, Scalability, and CAP(Consistency-Availability-Partition Tolerance).

Cost-performance is important when dealing with big data because the storage cost can be huge if using sql database to storage large amount of data. Dezyre mention that "RDBMS requires a higher degree of Normalization i.e. data needs to be broken down into several small logical tables to avoid data redundancy and duplication" [3], even normalization can meke database become well organized, but the cost of normalization will be huge if the data type and amount is huge. Nosql using collection to storage data which allow user using less time to import and search data, no need to classify the type of each file in nosql database.

The scalability of nosql database allow user improve their application performance and response speed, according to dezyre's idea "NoSQL Databases like the HBase, Couchbase and MongoD, scale horizontally with the addition ofextra nodes (commodity database servers) to the resource pool, so that the load can be distributed easily. " [3];the tables in sql database have relatedness, so we usually need to use join function to select data, so every data for one application must to be in the same server. The files in nosql have no relatedness, so storage the data in different server is feasible for nosql, then when the application growth require more servers to support, nosql database can simply adding servers to meet the business growth need.

Eric Brewer point out the cap theory for distributed system, "The Availability and Consistency that I mentioned comes, of course, from the misunderstood CAP theorem, that - so people say states that you can only choose 2 out of the 3 Consistency: every read would get you the most recent write Availability: every node (if not

failed) always executes queries Partition-tolerance: even if the connections between nodes are down, the other two (A & C) promises, are kept. " [7]. This theory represent that a distributed system can not satisfied consistency, availability, and partition-tolerance in the same time. For sql database,"ACID (an acronym for Atomicity, Consistency Isolation, Durability) is a concept that Database Professionals generally look for when evaluating databases and application architectures. For a reliable database all these four attributes should be achieved." [8], the ACID theory for SQL database is seems like making sql database more reliable than nosql database. But reliable is the contradiction of performance, which mean if the business become more complicated, the performance of sql database will be decrease.

3 THE ADVANTAGE OF USING NOSQL DATABASE FOR BIG DATA APPLICATIONS

There are three main advantages of using nosql database for big data application, flexible data model, high scalability, high performance. The data model in nosql database is flexible, user don't need to set up the file property at the beginning and custom the data storage formatting at anytime[5]. For example, if our big data analysis application is getting data from the third party platform like face book, the data from the third party platform is multifarious like target user information, physical and social graph. So the flexible data model allow application collect multifarious kind of data without bother the type of data which going to collect advance. When collecting new kind of data, nosql no need edit the table or create a new column which allow the developer focus more on the analysis area instead of modify the collecting data type.

High scalability can save huge cost for upgrade the database to meet the business need. The extend model for sql database is scale up, which mean if the user is increasing, we need to have better server for satisfy the new usage [6]. Better server need better CPU, hard disk, and ram to content all the table, this make the upgrade become super expensive. In the other side, the upgrade method for nosql is much economic. At previous section, I mention that the data storage in nosql is distributed, so the upgrade method for nosql database is horizontally expanded. For fit the increasing of data storage, we just need to add new server which has exactly same spec or even lower spec, because we don't need to put all the data into one server. Cassandra is one of the best example, the architecture of cassandra is similar to p2p model, so we can simply add new node to expend the cluster [2]. Therefore the scalability of nosql make upgrade much more easier and cheaper for big data application.

Nosql have really high performance in reading and update data because of the data model of nosql is simply and no relational. According to the Penchikala's article [9], Sql using Query Cache, so the cache will lose efficacy when every time update the cache, so this kind of cache have lowe performance; The Cache of nosql is recording level which have much higher performance.

4 CONCLUSIONS

The limitation of traditional sql database no longer can meet the demand of big data applicaiton, because the performance of sql database to deal large amount of data is lower. Under the mass data application environment, nosql has more agility to deploy new model on big data application, and the expand of nosql database is easier and cheap compare with sql database.

5 ACKNOWLEDGEMENT

I would like to take this chance to thanks to my tutor Miao, in process on reviewing my paper, he gave me many useful comments and advises. At the same time, I would like to thanks my instructor laszewsk, give me useful knowledge about how to write a report on Latex format. Finally, I would love to thanks my friends who working for Microsoft give me many idea about my nosql database.

REFERENCES

[1] Artur Ejsmont,. 2017. Insert performance comparison of NoSQL vs SQL servers. (2017). http://artur.ejsmont.org/blog/content/insert-performance-comparison-of-nosql-vs-sql-servers [Online; accessed 11-Nov-2017].

[2] Rick Cattell. 2010. Scalable SQL and NoSQL data stores. *In Pervasive Computing and Applications (ICPCA)* 2011 6th International Conference on, 363-366 (2010), 2.

[3] dezyre. 2017. NoSQL vs SQL- 4 Reasons Why NoSQL is better for Big Data applicationss. (2017). https://www.dezyre.com/article/nosql-vs-sql-4-reasons-why-nosql-is-better-for-big-data-applications/86 [Online; accessed 11-Nov-2017].

[4] Jepsen . 2017. MongoDB stale reads. (2017). https://aphyr.com/posts/322-call-me-maybe-mongodb-stale-reads [Online; accessed 11-Nov-2017].

[5] K. Orend. 2010. Analysis and Classification of NoSQL Databases and Evaluation of their Ability to Replace an Object-relational Persistence Layer. (2010). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.483&rep=rep1&type=pdf [Online; accessed 20-October-2017].

[6] A. Lakshman and P Malik. 2010. Cassandrafi!?A decentralized structured storage system. *Operating systems review*, 44(2), 35. (2010), 2.

[7] Lior Messinger . 2013. Better explaining the CAP Theorem. (2013). https://dzone.com/articles/better-explaining-cap-theorem [Online; accessed 11-Nov-2017].

[8] Pinal Dave. 2007. SQL SERVER fi?! ACID (Atomicity, Consistency, Isolation, Durability). (2007). https://blog.sqlauthority.com/2007/12/09/sql-server-acid-atomicity-consistency-isolation-durability/ [Online; accessed 11-Nov-2017].

[9] Srini Penchikala . 2017. Distributed Cache as a NoSQL Data Store? (2017). https://www.infoq.com/news/2011/11/distributed-cache-nosql-data-sto [Online; accessed 11-Nov-2017].

[10] Wiki. 2017. NoSQL. (2017). https://en.wikipedia.org/wiki/NoSQL [Online; accessed 11-Nov-2017].

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 firmfis systems have been getting better by the day and expected to be even more superior in the near future.

2 PRODUCT RECOMMENDER SYSTEM

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

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

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

3 BIG DATA FOR DYNAMIC PRICING

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

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

4 BIG DATA AND CUSTOMER SERVICE

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

5 ONE CLICK ORDERING

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

6 USING BIG DATA TO SUPPORT OTHER COMPANIES

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

7 BIG DATA TECHNOLOGIES

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

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

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

8 CONCLUSION

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

ACKNOWLEDGMENTS

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

REFERENCES

[1] Amazon. 2017. Big Data on AWS. (2017). https://aws.amzon.com/big-data/

[2] Roy F Baumeister. 2002. Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of consumer Research* 52, 4 (2002), 670–676.

[3] Marc L Berger & Vitalii Doban. 2014. Big data, advanced analytics and the future of comparative effectiveness research. *Journal of company effectiveness research* 3, 2 (2014), 167–176.

[4] Randal E. Bryant & Randy H. Katz & Edward D. Lazowska. 2008. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. (2008). https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf

[5] Benjamin Reed Shiller. 2014. First-Degree Price Discrimination Using Big Data. (2014). http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data_Jan_18,_2014.pdf

[6] Hsinchun Chen & Roger H L Chiang & Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012), 1165–1188.

Using MQTT for Communication in IoT Applications

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

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

ABSTRACT

With the increase in the number of edge devices and their applications in real world applications such as sensor networks, it is crucial to enable fast communication between the sensing devices and actuators, which may not be directly connected. To allow services that are built on different software and hardware platforms 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. We have evaluated the feasibility of using MQTT with the help of sentient architecture inspired dendrites.

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 an increased need to allow these devices to communicate quickly and securely. In many cases these edge devices have very limited memory and need to conserve power. The computing power on some of these devices is so limited that the sensory data need to be analyzed remotely. Furthermore, they may not even have enough computing capacity to process traditional HTTP web requests efficiently [2, 10] or these traditional Web-based services are too resource hungry. 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 circuits or embedded chips on edge computing sensors [2].

Message Queue Telemetry Transport (MQTT) is a lightweight machine to machine (M2M) messaging protocol, based on a client/server publish-subscribe model. It provides an elegant solution for such scenarios.

WHICH 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. However du to leveraging TCP/IP it is reliable. 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].

The current support for MQTT is conducted by the Eclipse Paho project [5].

As MQTT is a protocol many different clients exist in various languages. This includes languages such as C, Python and Lua.

Common brokers include the open source Mosquitto broker [15] and a Really Small Message Broker from IBM [5].

what is the real small broker, this seems unclear

.

2 MQTT DETAILS

MQTT works via a publish-subscribe model that contains 3 entities: (1) a publisher, that sends a message, (2) a broker, that maintains queue of all messages based on topics and (3) multiple subscribers that subscribe to various topics they are interested in [17].

This allows for decoupling of functionality at various levels. The publisher and subscriber do not need to be close to each other and do not need to know each others identity. They need only to know the broker, as the publisher and the subscribers do not have to be running either at the same time nor on the same hardware [12].

2.1 Topics

MQTT implements a hierarchy of topics that are relates to all messages. These topics are recognised by strings separated by a forward-slash (/), where each part represents a different topic level. This is a common model introduced in file systems but also in internet URLs.

A topic looks therefore as follows: *topic-level0/topic-level1/topic-level2*.

All subscribers subscribe to different topics via the broker. Subscribing to *topic-level0* allows the subscriber to receive all messages that are associated with topics that start with *topic-level0*. This 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 messaged to the subscribers, but it does not provide any form of security by default to make is useful for resource constrained IoT devices. “It allows the use of username and password for authentication, but by default this information is sent as plain text over the network, making it susceptible to man-in-the middle attacks” [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 a the identity of the server and client with the help of a handshake mechanism which uses trust certificates to establish identities before encrypted communication can take place [4]. If the handshake is not completed for some reason, the connection is not established and no messages are exchanged [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 short-lived then most of the time can be spent in the handshake itself, which may take up few kilobytes of bandwidth. In case the connections are short-lived, temporary session IDs and session tickets can be used to resume a session instead of repeating the handshake process. If the connections are long term, the overhead of the handshake is negligible and TLS/SSL security should be used [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 ans the user and are signed by a trusted authorization server [9].

When connecting to the broker this token can be used to check whether the client is authorised to connect at this time or not. Additionally the same validations can be used when publishing or subscribing to the broker. The broker may use a third party resource such as LDAP (lightweight directory access protocol) to look up authorizations for the client [9]. Since there can be a large number of clients and it can become impractical to authorize everyone, clients may be grouped and the authorizations may be checked for each group [4].

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

MQTT broker services can be utilized for enterprise and production environments. EMQ (Erlang MQTT Broker) provides a highly scalable, distributed and reliable MQTT broker that can be used in enterprise-grade applications [8].

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 raspbean 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 ot 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*) 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 incude 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 ans 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

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. (augist 2017). http://blog.xively.com/why-mqtt-is-the-protocol-of-choice-for-the-iot/

[3] cloudmesh. 2017. cloudmesh.pi. github. (ocober 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

[5] eclipse. [n. d.]. mqtt broker. eclipse mosquitto website. ([n. d.]). https://mosquitto.org/

[6] eclipse paho. [n. d.]. Python Client - documentation. eclipse paho wensite. ([n. d.]). https://www.eclipse.org/paho/clients/python/docs/

[7] elastic.io. [n. d.]. ELK stack. elastic.io website. ([n. d.]). https://www.elastic.co/products

[8] erlang mqtt. [n. d.]. erlang mqtt broker. wmqtt website. ([n. d.]). http://emqtt.io/docs/v2/index.html

[9] hive mq. [n. d.]. MQTT Security Fundamentals: OAuth 2.0 & MQTT. hivemq website. ([n. d.]). https://www.hivemq.com/blog/mqtt-security-fundamentals-oauth-2-0-mqtt

[10] hivemq. [n. d.]. intrwebsite mqtt. hivemq website. ([n. d.]). https://www.hivemq.com/blog/mqtt-essentials-part-1-introducing-mqtt

[11] Dexter Industries. 2017. Grovepi. Dexteer Industries website. (2017). https://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.seeed.cc/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.seeed.cc/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=808771136 [Online; accessed 6-November-2017].

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 ant 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 is 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.

Big Data and Artificial Intelligence with Computer Vision

Bharat Mallala
Indiana University
Bloomington, IN 47408, USA
bmallala@iu.edu

ABSTRACT

Big data refers to a problem of dealing with huge volumes of data. With the increase in the amount of data generated every day from various fields, it is becoming extremely hard to store and process this data efficiently. Artificial Intelligence is a research field aiming at replicating human work through machines. Computer vision refers to a research area within AI dealing with training computers to recognize certain subjects of interest. With the exponential growth of AI and computer vision in the recent years, there is need to address the big data problem associated with it.

KEYWORDS

hid215, Artificial Intelligence, Computer vision, Perceptron Deep Learning, Convolutional Neural Networks

1 INTRODUCTION

Artificial Intelligence is an aim at replicating human intelligence through machines. The term AI was in existence form the late 1950's during which there was a lot of enthusiasm on its potential during which Alan Turing introduced the Turing test. A lot of research was carried out in AI during the 1960's with the introduction of perceptron theory and its ability to solve problems. There was a major setback for AI in the early 1970's during which Minsky in his book on perceptrons has pointed out the major drawbacks of perceptrons in dealing with complex problems.[2]

There has been a consistent growth in AI from the 1990's with the introduction of the statistical approach to problem-solving. With the increase in the use of Big data from 2010, there was a lot of development in the field of AI with many voice assistants, self-driving cars, automated robots etc. Many AI problems which were np-hard previously would now take minutes to solve thanks to recent advancements in big data. Professor Crandall quoted during his lecture quoted that "Problems that seem to require intelligence usually require exploring multiple choices".[2], which is a way of exhibiting intelligence without actually having it. For example for a machine to win a tic-tac-toe game against a human it basically has to explore all the possible choices, it can make from any given state.

Hence we can map an AI problem as a search problem, but it usually requires searching through huge search spaces, this is when it becomes a Big data problem. For example, for a computer t win against a human in chess it needs to search through hundreds of thousands of states. To deal with such huge data, there is a need to apply big data technologies to better store data and efficiently manage it. Ai problems typically involve using both structured and unstructured kind of data in huge volumes. The traditional RDBMS methods have a hard time dealing with unstructured data. This is an area where Big data shines with the ability to deal with both structured and unstructured data efficiently.

"Computer vision is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos".[5] It is an area of research within AI that aims at recognizing subjects in an environment from images or videos. Convolutional Neural Networks shines at image classification from images with minimal prepossessing of the input variables and still manages to obtain better classification. With the exponential rise of AI and especially CNN lately, there is an increased interest in computer vision for researchers. Computer vision involves training the model with huge sets of images and videos which indeed needs to be addressed using big data technologies.

2 RETHINKING THE INCEPTION ARCHITECTURE FOR COMPUTER VISION

2.1 CNN's for Computer vision

Convolutional Neural Networks(CNN's) is the key to advancements in computer vision in the recent years with its low parameter count and computational efficacy. A CNN has multiple layers with perceptrons in them that help in the information flow from the input to output. A CNN typically has filters that are mapped across the original image to extract useful features from the image. During the training phase of the model, CNN learns the values of the filters. A CNN architecture has mainly two phases convolution phase and pooling phase. In the convolution phase features are extracted from the image using filters. In the pooling phase, the width of the feature map is reduced by applying various techniques. This is done to remove unnecessary data from the features. These stages are iterated till the desired features are obtained from the image [6]. Figure 1 shows a typical CNN architecture [3].

2.2 Design

A lot of design principles need to be followed when designing a CNN. With the numerous number of iterations required for the CNN phases, a good design decision can vastly improve the efficiency.Avoiding bottlenecks/cycles in the CNN helps in smooth flow of information which otherwise can drastically reduce the performance of the CNN. The CNN is usually trained locally within the network. Using a high dimensional representation of the feature space help to in training the network at a relatively faster rate. Pooling/Spacial aggregation of the CNN helps to reduce the dimensions of the feature space which also helps in faster training. This is most helpful when we are dealing with big data and we have a huge set of training images. When dealing with huge CNN with huge depth and width of the network, there is a need to strike a balance among width and depth of the image. The depth of CNN

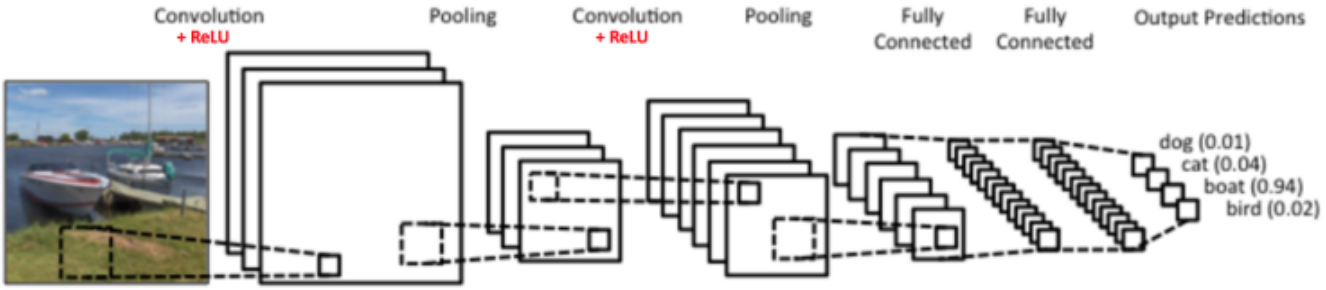


Figure 1: Convolutional Network Architecture [3]

is determined by the size of the filter used at each iteration of the network. Optimality can be obtained when increasing each of them in parallel for every iteration.[1]

2.3 Method

Once the training set of images is obtained, the depth of the CNN should be formulated i.e. 1x1 layer of 3x3 layer etc. Now that the depth is determined, the next step is to determine the size of the filters. CNN with large filters usually tends to be expensive when it comes to computational cost. Hence the ideal approach will be to start with a reasonable filter size and gradually reduce it in every iteration. But a very small filter may not extract enough features from the image, hence it is important to reduce the filter size and also making sure enough features are extracted.[1]

2.4 Role of Big data

With the exponential increase in the training sample for the CNN, the depth of the CNN would drastically increase making it not computationally feasible. Big data techniques needs to applied to CNN to meet the vast training requirements of computer vision. The training sample is stored across multiple node of the Hadoop architecture making more easily accessible. Each stage of the CNN is carried across multiple nodes making it computational feasible. Since Hadoop also supports semi and unstructured data as well, unprocessed images can also be trained with ease using CNN.[1]

3 ARTIFICIAL INTELLIGENCE AND BIG DATA

3.1 Contributions of AI

With ever increasing data in the field of AI, researchers are outsourcing some of the AI tasks such as pattern recognition, deep learning to other parallel computing based methods. Typical AI systems with their extensive use in decision making for major businesses around the globe, need to make decisions with a specific time limit. As formulated earlier if we map an AI problem as a search problem, searching through huge spaces in relatively less amount of time is a cumbersome task to do. For example, while playing a game of chess against a human, a computer cannot take an infinite amount of time to make the next move as it searches across all the possible states. Also, most AI applications typically deal with structured data as their input. But with the increase in the amount of data AI needs to be processed currently the approach to deal only with structured data no longer works. There is a need to apply other techniques to process the unstructured and semi-structured data and use it for training the AI models.[4]

3.2 Limitations of AI algorithms

Till date, much of the research is centered on using AI on a single machine which has its limitations when it comes to data storage and computational efficiency. Many of the existing data mining algorithms are limited when it comes to dealing with big data. Data might consist of some inconsistencies, for example missing data, incorrect data, etc. which leads to a majority of the AI algorithms to obtain a good accuracy[4].

3.3 Mapreduce Approach

Using MapReduce and address most of the above-stated limitations of AI algorithms. MapReduce in the recent times has been used for parallel-processing which can be quite useful for AI algorithms. With the approach of using parallel processing many of the machine learning algorithms, we can train the model to learn simultaneously from multiple machines, thereby drastically reducing the overall computational cost and time. Researchers have also developed a machine learning module on Hadoop called Mahout. Mahout is responsible to run all of the tasks that a traditional system can do but within a relatively less amount of time. This done using parallel data storage and processing feature of Hadoop.[4]

Typically in a Hadoop environment, any AI problem is divided into numerous subproblems depending upon the size of the cluster. The training data similarly is subdivided according to the tasks. Now, these tasks are allocated to all the data nodes. The metadata corresponding to the data nodes will be stored in the name-node. There will also be a secondary name-node in case the name-node fails. The zookeeper component is responsible to allocate tasks to data nodes. All these work together in real time to obtain parallel processing of the data. [4]

3.4 Issues with AI and Big data

AI and big data complement each other under numerous occasions. But there are a lot of issues when it comes to compatibility between AI and Big data. Firstly many of the iterative generic Machine learning algorithms are difficult to be used in the Hadoop environment. These algorithms do not come included with the Mahout module.

So researchers are trying to address this issue of compatibility. Secondly, there has been an inconsistency in Data visualization part of AI. AI algorithms, when used in traditional systems, are capable of creating some great visualizations to make the results of the model more appearing. By using AI within Hadoop environment the level of visualizations supported on Hadoop is mediocre. Third, AI algorithms can be used to model real-time data and make inferences from the same. When it comes to Hadoop we have a component 'Flume', But it does not support the level of detail as the traditional systems.Fourth, Hadoop environment is not so well equipped when it comes to dealing with audio/video file which the traditional algorithms are more capable of. [4]

4 CONCLUSION

Artificial Intelligence with its ability to take over the world has some limitations when it comes with Big data, i.e. huge volumes and different varieties of data. With the advancements in Big data and its technologies, this gap between the two can be reasonably reduced by making modifications to existing AI approaches to make it compatible with Big data. Mahout is one of the Big data modules by which we can apply many of the AI algorithms to Big data. Advancements in Computer vision has led to extensive use of CNN's for applications such as image recognition. Big data has provided a means to apply computer vision techniques on a much larger scale with low computational cost in a relatively lesser time.

ACKNOWLEDGMENTS

I would like to thank Dr. Gregor von Laszewski and the AI's for all the help they have provided for this paper.

REFERENCES

[1] Sergey Ioffe Jon Shlens Christian Szegedy, Vincent Vanhouc and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 22, 21 (2016), 2818–2826. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html

[2] David Crandall. 2017. Lecture on Artificial Intelligence. (2017). https://iu.instructure.com/courses/1649672/files/72178167?module_item_id=16172128

[3] Google. 2017. "Google images". (2017). https://www.google.com/search?q=convolutional+neural+network&num=20&newwindow=1&rlz=1C1CHBF_enUS759US759&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi9tsLMwqjXAhWa14MKHb1HCEgQ_AUICigB&biw=1707&bih=826&dpr=1.13#imgrc=dBgDZd3GJF-P5M:

[4] D. E. O'Leary. 2013. Artificial Intelligence and Big Data. *IEEE* 28 (2013), 96 – 99. https://doi.org/10.1109/MIS.2013.39

[5] Wikipedia. 2016. "Wiki website". (2016). https://en.wikipedia.org/wiki/Computer_vision

[6] Donald Williamson. 2016. Lecture on Convolutional Neural Networks. (2016). https://iu.instructure.com/courses/1600135/files/folder/Lecture%20Slides?preview=69655358

Big Data With Apache Spark

YuanMingHuang
Indiana Univeristy Bloominton
Bloomington, IN 47408
huang226@indiana.edu

ABSTRACT

This study provided a short overview of big data processing framework Apache Spark, and mainly introduce Apache Spark frame, and compared to some other big data technologies like Hadoop. Then, introduced what is new in Apache Spark and Spark Ecosystem. It will help people to better learn about Apache Spark.

KEYWORDS

Big Data, Framework, Hadoop, Apache Spark

1 INTRODUCTION

Apache Spark is an open source big data processing framework. However, it is built around speed. The advantage of it is easy to use and sophisticated analytics[3**?**]. It was originally developed in 2009 in UC Berkeleyfis AMPLab, and open sourced in 2010 as an Apache project. Comparing to other big data technologies, for example, Hadoop and Storm, it has the following advantages: Firstly, Spark provides users a more comprehensive, unified framework to manage big data processing requirements with a great number of data sets that are diverse in nature as well as the source of data. Secondly, Spark makes it possible for applications in Hadoop clusters to run up to around 100 times faster in memory and 10 times faster even when it is running on disk. Thirdly, Spark allows users quickly write applications in some different programming languages, such as Java, Scala, or Python. At the meanwhile, it comes with a built-in set of over 80 high-level operators. As well as you can use it interactively to query data within the shell. In addition to Map and Reduce operations, it can also support SQL queries, streaming data, machine learning and graph data processing. Thus, Developers can use these abilities stand-alone or combine them to run in an individual data pipeline use case. When you are first installment of Apache Spark article series, users will look at what Spark is, how it compares with a typical MapReduce solution and how it provides a complete suite of tools for big data processing.

2 WHY IT IS BETTER THAN HADOOP

Hadoop is a big data processing technology. It has been around for 10 years and has proven to be the solution of choice for processing large data sets. MapReduce is a very powerful solution for those one-pass computations, however, sometimes it not that efficient for use cases if it requires multi-pass computations and algorithms. Every step in the data processing workflow will have one Map phase and one Reduce phase , thus users will need to convert use case into MapReduce pattern ,and then it will be able to leverage this solution. Hence, if we want to do something complicated, we would have to string together a series of MapReduce jobs and execute them in sequence. Each step of those jobs was time killer, and none of them could start unless the previous job had already finished completely. However, Spark will allow programmers to develop complex, multi-step data pipelines using directed acyclic graph pattern. It makes it easier and will save a great number of time. In the meanwhile, it also supports in-memory data sharing across DAGs, so that even though it is some different jobs, it can also work with the same data which makes it easier to operate and control.

After those analysis, we will find that we can see Spark as an alternative to Hadoop MapReduce rather than a replacement to Hadoop. Itfis not intended to replace Hadoop but to provide a more comprehensive and unified [4**?**] solution to manage and run different big data use cases and requirements.

3 WHAT IS IN SPARK

With shuffles in the data processing, Spark takes MapReduce to the next level. And With capabilities like in-memory data storage and near real-time processing, it improved its performance[1**?**], so it can be several times faster than other big data technologies. Within Spark, it also provides a higher-level API to improve developer productivity and a consistent architect model for big data solutions. And Spark also supports lazy evaluation of big data queries, it will help with optimization of the steps in data processing workflows.

Apart from above, Spark also include more functions to support more than just Map and Reduce, arbitrary operator graphs and a more interactive shell for Scala and Python. Because Spark is written in Scala Programming Language and runs on Java Virtual Machine environment. So, it can support the following languages for developing applications using Spark, like Scala, Java, Python, Clojure and R.

4 SPARK ECOSYSTEM

In addition to Spark Core API, there still are some other additional libraries. All of them are part of the Spark ecosystem, and they provide additional capabilities in Big Data analytics and Machine Learning areas. These libraries are:

4.1 Spark Streaming

Spark Streaming can be used for processing the real-time streaming data. It is based on micro batch style of computing and processing and it uses the DStream which is basically a series of RDDs to process the real-time data.

4.2 Spark SQL

Spark SQL [2**?**] provides the ability to expose the Spark datasets over JDBC API and it allows running the SQL like queries on Spark data using traditional BI and visualization tools. Spark SQL allows the users to ETL their data from different formats.

4.3 Spark MLlib

MLlib is Sparkfis scalable machine learning library consisting of common learning algorithms and utilities, it includes classification, regression, clustering, collaborative filtering, dimensionality reduction, and underlying optimization primitives.

4.4 Spark GraphX

GraphX is the new Spark API for graphs and graph-parallel computation. At a high level, GraphX extends the Spark RDD by introducing the Resilient Distributed Property Graph. In order to support graph computation, GraphX exposes a set of fundamental operators like subgraph, joinVertices, and aggregateMessages and optimized variant of the Pregel API.

5 A RESILIENT STRUCTURE

Resilient Distributed Dataset is the core concept in Spark framework. We can consider about RDD as a table in a database. And it can hold a great numbers types of data. And those data will be stored on different partitions. They help to rearrange the computations and optimize the data processing. They are also fault tolerance because RDDs knowhow to recreate and compute the datasets. In the meanwhile, RDDs are immutable, so we can modify an RDD with a transformation. And then the transformation will return us a new RDD whereas the original RDD remains the same.

6 CONCLUSION

Apache Spark as an open source big data processing framework. it has a very fast speed than other technologies. In the meanwhile, it provides a lot of tools to process data, it provides a new way to improve the efficiency in processing dataset. And we can use different programming languages to finish different job within Spark. It has many advantages than Hadoop, but it not born for replacing Hadoop, but helping to improve the capability of processing data and speed to finish big data jobs.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Michael Armbrust, Tathagata Das, Aaron Davidson, Ali Ghodsi, Andrew Or, Josh Rosen, Ion Stoica, Patrick Wendell, Reynold Xin, and Matei Zaharia. 2015. Scaling Spark in the Real World: Performance and Usability. *Proc. VLDB Endow.* 8, 12 (Aug. 2015), 1840–1843. https://doi.org/10.14778/2824032.2824080

[2] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. 2015. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*. ACM, New York, NY, USA, 1383–1394. https://doi.org/10.1145/2723372.2742797

[3] Shyam R., Bharathi Ganesh H.B., Sachin Kumar S., Prabaharan Poornachandran, and Soman K.P. 2015. Apache Spark a Big Data Analytics Platform for Smart Grid. *Procedia Technology* 21, Supplement C (2015), 171 – 178. https://doi.org/10.1016/j.protcy.2015.10.085 SMART GRID TECHNOLOGIES.

[4] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM* 59, 11 (Oct. 2016), 56–65. https://doi.org/10.1145/2934664

Big Data Applications in Virtual Assistants

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

ABSTRACT

In the age of big data, artificial intelligence and speech recognition techniques have been widely used in numerous big data technologies and applications. Among those are virtual assistants which could potentially lead to the future evolution of big data. We list various virtual assistants currently in the industry developed by giants such as Google, Microsoft, Amazon and Apple. We then follow up by discussing some future development of virtual assistants.

KEYWORDS

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

1 INTRODUCTION

Since the early 2000s, big data has been a popular word among the tech industries. It is a term that describes data sets which are so huge that normal data management techniques are not enough to process them. Nowadays, big data usually refers to data science analytics methods such as predictive modeling, machine learning, data mining, speech recognition and so on [5]. Big data is typically described by the 4Vs, that is Volume, Velocity, Variety and veracity. Volume stands for the size of the data, velocity refers to the data processing speed, variety means different types of data and finally veracity represents the quality of the data.

As the name suggests, big data is big. The size of the data could go up to petabytes or even exabytes which is 1000 petabytes. Previously, to process this amount of data is essentially impossible. However, as the year goes by, the ability to analyze big data has increased significantly thanks to improved mathematical algorithms and powerful computers [5].

In the era of big data, the opportunities are immense. Big data not only will help all kinds of companies and organizations to make better services and decisions but also will aid us in everyday tasks such as writing emails, sending texts, shopping online, listening to music, watching TV, traveling around the globe, managing schedules, and so on [5]. This led to to the birth of virtual assistants which are software that can perform various tasks described above for individuals. The major technologies behind virtual assistants are artificial intelligence and speech recognition which would not be possible without the use of big data [1]. At present, more and more companies and organizations are beginning to employ the use of virtual assistants which utilize artificial intelligence. As a result, along with the help of machine learning algorithms, more and more applications, technologies and systems will be able to obtain information on their own such as user behaviors to predict and provide useful suggestions to individuals with more personalized experiences [3].

Virtual assistants work usually via three different methods: by texts, or voices or sometimes images. A virtual assistant may work via more than one method such as Goggle Assistant which can understand texts and voices. After users input their commands either via texts or voices, the virtual assistants process those information received using a method called natural language processing which then translates user inputs into executable commands they can interpret.

2 CURRENT APPLICATIONS IN THE INDUSTRY

In this new era of virtual assistants, four major tech companies come into play: Google with Google assistant and Google Now, Apple with Siri, Microsoft with Cortana and Amazon with Alexa. All these virtual assistants have been successful in the market but the technologies behind them are still immature. The four giants have been competing with each other in this area for years and each of them wants to bring their own unique virtual assistant to the market.

Google has a variety of services such as its famous search engine Google, Gmail, Google maps, Google drive, etc. and signing up for a Google account gives you the ability to use all these services any time from any internet devices such as phones, tablets and laptops. Everything you do on these services is registered and kept in a personal cloud on Google's servers. With these information, Google assistant can learn your behaviours and styles in order to provide you with useful suggestions and recommendations before you even know it [5].

Compared to other companies, Google has a competitive edge in machine learning because Google has access to vast amount of information and resources from its popular search engine and cloud services such as Gmail, Google drive, Google docs, etc. In addition, Google also is connected with more than 1.5 billion smart-phone users on Android platform. These people are more likely to use Google's new virtual assistant systems in the future to manage their everyday tasks [5].

During its original release with the IPhone 4S back in 2011, Siri was a big surprise to the general public and it became a huge success as an early virtual assistant in the field. However, the initial release of Siri also received some negative reviews for its lack of information to give directions to nearby places as well as for its bad speech recognition to understand some certain English accents. Since then, Apple has been making progressive improvements on Siri in order to overcome these disadvantages it had before [6]. With information and data collected from Apple users all over the world, either their IPhone or laptops or desktops, Apple can provide

their users with a more advanced Siri to help their everyday lives [5].

In 2015, along with Microsoft's release of its newest operating system Windows 10 was the introduction of Cortana, the Microsoft version of a virtual assistant, finally entering the competition with Google and Apple. Furthermore, Microsoft provided the option for Windows users to download a Cortana app on their smart-phones so that they can share information across platforms (phones and computers) to manage their tasks on the go [5].

The difference between Microsoft and its competitors is that Microsoft has more experience in business software such as Microsoft Power BI. As a result, Cortana is integrated with various Microsoft apps like Skype and services to aid users in making better business decisions with ease. Cortana uses reminders to schedule your meetings, create to-do lists and so on. For example, you can create a reminder like, *Schedule a meeting with Tom at 10am*, in an app called Smart Sticky Notes either by telling Cortana with voice command or typing it in. Then Cortana will add that reminder to your list while keeping track of it for you. In addition, Microsoft is trying to add features to Cortana such as synchronizing reminders across different devices. This could potentially lead to the expansion of Cortana to create reminders and to-do lists with data from Office 365 [4].

As Google Assistant, Siri and Cortana dominated the virtual assistant market and continued to get smarter, Amazon - the world's largest Internet-based retailer, wanted a piece of the action as well. In 2014, Amazon entered the competition with its Amazon Alexa which was capable of voice integration, providing real time traffic, weather and news information, playing music, setting alarms and timers as well as making to-do-lists and reminders. Alexa can also be used as a smart home system where it can control smart home devices such as lights and thermostats [4].

What separates Amazon with its competitors is that in 2015 Amazon released the Alexa Skills Kit which allowed designers and developers to build their own apps and skills through the technologies behind Alexa and integrate them in any Alexa-enabled devices. The Alexa Skills Kit received highly critical claims and subsequently, companies and organizations now have a wide range of apps and skills to choose from for their own business needs. Google and Microsoft soon pursued Amazon in its path as well with the release of Google and Cortana Skills [4].

Eventually, in the future, people will only want one virtual assistant which is available anywhere you go and can do various tasks you request. However, with the current trend in the industry, we will mostly likely to see different virtual assistants specialize in their own unique way. For example, Google Assistant will excel in giving driving directions and gathering information while Microsoft Cortana will focus on providing people with a powerful and enjoyable gaming experience. Alexa will specialize in bringing well personalized recommendations for shopping [1].

3 THE FUTURE OF VIRTUAL ASSISTANTS

The possibility for the future of virtual assistants are vast. For shopping, the next generation virtual assistants should be able to provide recommendations with a focus on direct customer interactions and also use locations to create a smooth and more localized shopping experiences [3]. In auto-mobile industry, virtual assistants are starting to integrate themselves with smart vehicles. As a fact, Microsoft already announced its *connected-vehicle* platform this year which tries to expand its Cortana capabilities on smart cars [1]. Nissan revealed in January this year that it planned to use Microsoft's vehicle platform in its cars. In addition, Apple has established its Car-Play system on Siri-enabled smart cars or via Siri on smart-phones. Furthermore, Hyundai is putting Google Assistant and Amazon Alexa into their cars to create some cool functions such as starting your car from the living room [1].

Amy, Shae and Otto, these are three new virtual assistants on the rise in the market. Though they are only in either beta test or prototype stage, they represent the future of virtual assistants and what we can achieve [2].

Amy is created by a start-up company in New York called *x.ai*. In its current stage, Amy is only an email address: *amy@x.ai*. The goal for *x.ai* is to integrate Amy with other platforms such as Amazon Alexa, Slack and so on. Amy is invisible meaning there is nothing you need to install. For now, Amy is good at scheduling meetings. For example, when you want to schedule a meeting, simply cc the details to Amy's email address and your meeting will be scheduled. Amy can understand our daily language because it is powered by natural language processing techniques. For example, When you want to schedule a meeting with a colleague, you can send him an email and cc Amy, saying something like *Hi, how about we meet up some time*. Amy will then look into your calendar and come up with a time to meet.

Amy can also interact with others. For example, when the person you plan to meet replies with certain conditions, Amy can take those conditions into account and suggest another meeting time all by herself without any human interventions. However, you can always ask Amy progress of the meeting scheduling by sending her an email. Last but not least, Amy can also reschedule meetings in the same way. The only thing you need to do is to send an email to Amy saying something like *Cancel and reschedule my meetings next week*. No matter how many meetings you have, Amy will contact all those people to reschedule the meetings and then update your calendar accordingly [2].

Shae is created by a company called *Personal Health 360* or *PH 360* to help people live a fresh and healthy life everyday by recommending health related suggestions and information. The company says that it uses big data as in more than ten thousand data points and several hundred mathematical algorithms to provide personalized health advice to users. The company claims these algorithms take into account factors such as medical history, family history, body type, demographics and even location data such as weather. These data are collected via surveys given by the company when users register for their accounts. If you have wearables like Fitbit, Shae can also obtain bio-metrics information from it to detect your stress level. In cases where it detects high level of stress, Shae will ask you whether you are experiencing any discomforts such as heart attacks. Like other health apps, Shae will give you advice on how to have a healthy diet and when to exercise as well as keeping an eye on your body measurements [2].

Otto is develop by Samsung to be a home virtual assistant like the Amazon Echo. Like the Echo, Otto is Internet-enabled and can interact with users through voice commands. It can also control

smart devices in your home such as lights and thermostats. In addition, Otto can play music and help you in shopping online such as finding or placing an order. However, the difference between Otto and Echo is that Otto can also act as a security camera for your home when you are away. It can stream live videos to your personal smart-phone devices and you can also remotely control Otto's camera to look around the room [2].

All these three virtual assistants are unique in their own ways with Amy specializing in scheduling meetings, Shae in providing information on healthy lives and Otto in making homes safer. They demonstrate how smart, powerful and helpful virtual assistants will be in the future [2].

4 CONCLUSION

Virtual assistants are incredibly helpful artificial intelligence machines which utilize machine learning and speech recognition techniques to learn preferences and behaviours to make our everyday lives better. We give a brief introduction to virtual assistants and how they work. In addition, we list several current virtual assistant applications in the industry created by giants like Google, Apple, Microsoft and Amazon as well as how they differ in their specializations and unique usages. We further discuss the future of virtual assistants with emerging examples such as Amy, Shae and Otto to show how helpful virtual assistants can be in the future.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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 Haddop 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. 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 [4]. Figure 1 shows HDFS architecture.

HDFS is based on master/slave architecture where NameNode is the master server and DataNodes are the slave nodes. There can be

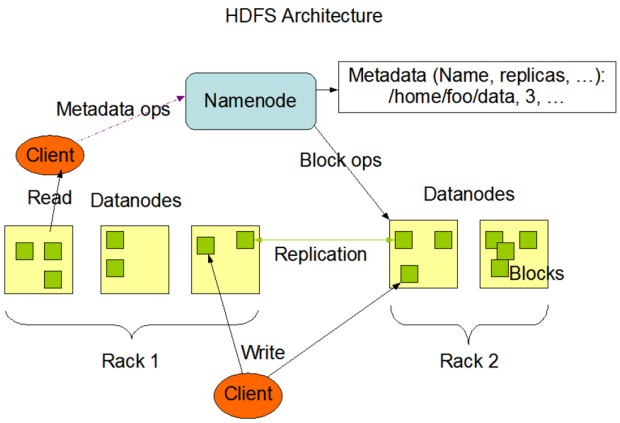


Figure 1: HDFS Architecture [4]

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.

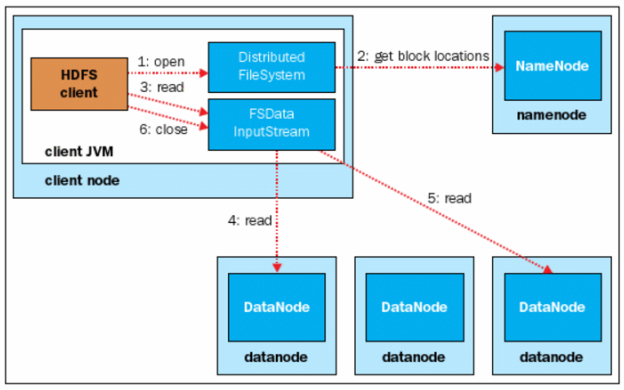


Figure 2: HDFS Read [1, Ch. 3, p. 38]

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

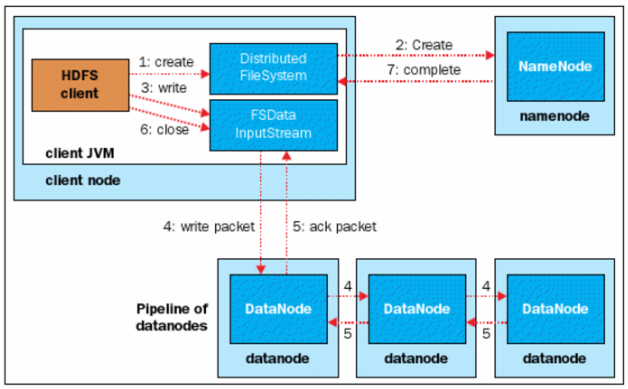


Figure 3: HDFS Write [1, Ch. 3, p. 39]

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.

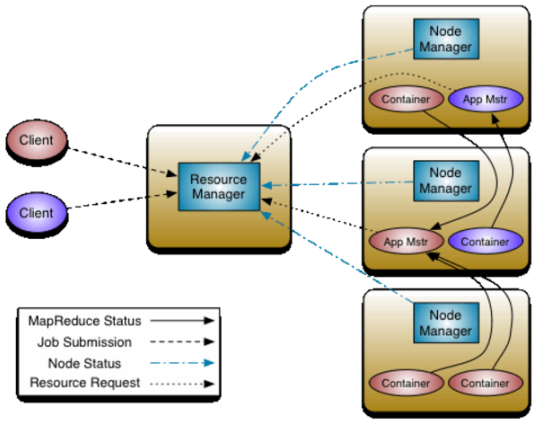


Figure 4: YARN Architecture [3]

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.

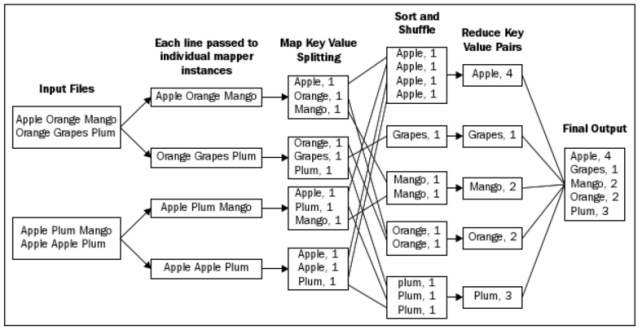


Figure 5: MapReduce Example [1, Ch. 3, p. 48]

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.

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].

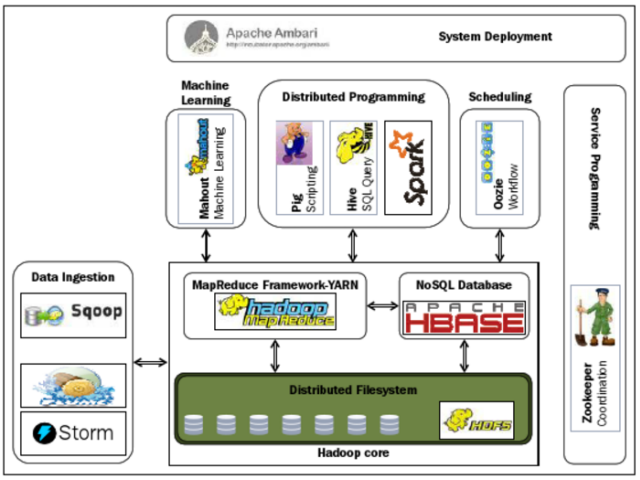


Figure 6: Hadoop Ecosystem [1, Ch. 2, p. 26]

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. MongoDB provides powerful query language,

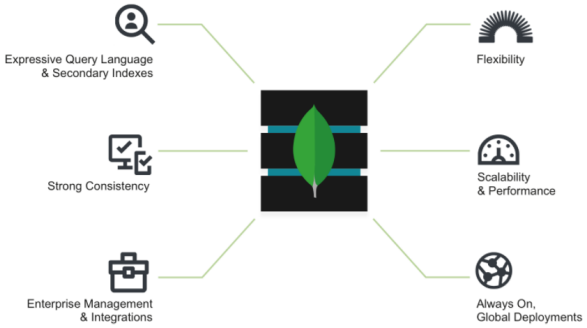


Figure 7: MongoDB Architecture [7]

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. It provides "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
[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

Big Data in Deep Space Telemetry and Navigation

Rick Carmickle
Indiana University School of Informatics and Computing
901 E 10th St
Bloomington, Indiana 47408
rcarmick@umail.iu.edu

ABSTRACT

Big data has begun to impact telemetry and navigation for manned and unmanned space travel. Major space programs are the entities which control the infrastructure of telemetry and deep space data. The current bottleneck in the volume, variety, and veracity of deep space and telemetry data is the transmission technology uses to communicate across space. Space Programs are poised to shift from radio-wave to laser and infrared communication methods which will allow greater volumes of data to be transmitted and quickly make big data more applicable to this type of data.

KEYWORDS

Big Data, i523, Telemetry, Deep Space, NASA Deep Space Network, Laser Communications Relay Demonstration

1 INTRODUCTION

Big data in space telemetry and navigation concerns the process of transmitting data between spacecraft and controllers on earth, and storing data from spacecraft. These Spacecraft can be either manned or unmanned and return different volumes and types of data depending on their purpose. The nearest spacecraft are those in low-earth orbit [1], between 400 and 1000 miles above the earth's surface and the farthest is Voyager 1 which is currently just over 140 AU (Astronomical Units) or approximately 13 billion miles from Earth [10]. The major space communications networks are operated by the United States, Europe, Russia, China, India, and Japan. Communications between spacecraft, satellites, and earth-bound communication centers has historically been facilitated with radio and microwave wavelength signals [16]. These frequencies allow communication at speeds up to 100 megabits per second in the case of the Lunar Reconnaissance Orbiter [5]. Deep Space communications are, in late 2017, on the cusp of developments which will open the big data opportunities substantially [9]. The volumes of telemetry and science data transmitted between tracking stations and spacecraft are measured in megabits per second for unmanned spacecraft to gigabits per second for manned missions [5].

2 DEEP SPACE DATA INFRASTRUCTURE

Deep space and telemetry data, gathered at a rate which could be considered big data, is done by the world's largest government space programs. The United States, Russia, Europe, China, Japan, and India have programs with ground tracking stations capable of receiving deep space data and launch capabilities to carry manned spacecraft to orbit or unmanned spacecraft beyond orbit [7].

2.1 NASA Deep Space Network

The NASA Deep Space Network (NASA DSN)has three core stations located in Goldstone in the Mjoave Desert, California, near Madrid, Spain, and near Canberra, Australia. Each location includes a 70 meter main antenna, a 34 meter high efficiency antenna, ten additional deep-space stations with parabolic reflector antennas, a 34 meter beam waveguide antenna which can incorporate new electronics easily, and a 26 meter antenna for tracking satellites in earth orbit [7]. NASA DSN is able to receive data with signal strength as weak "20 billion times weaker than the power level in a modern digital wristwatch battery" [4] which is the signal strength currently received from the Voyager I [10]. NASA DSN is equipped to communicate in S-band and X-band radio frequencies, near-infrared frequencies, and is equipped to be adapted to new data transmission technologies [4].

2.2 ESTRACK

The European Space Agency's (ESA) Deep Space Network [3] has nine core stations in South America, Europe, Australia, and Atlantic islands. ESTRACK has cooperative network stations through the NASA network, and in Canada, South Africa, and Japan. The sensitivity of ESTRACK rivals that of the NASA network and is also equipped to adapt to emerging technology beyond S- and X-band radio frequency [3].

2.3 Soviet Deep Space Network

The Russian State Corporation for Space Activities, Roscosmos, still carries out deep space communication with the Soviet Deep Space Network [14]. The three main antennas date to before the collapse of the Soviet Union, and have seen few hardware upgrades since. The network includes two 70 meter antennas, one in Crimea and one on the Eastern border of Russia along with a 64 meter antenna near Moscow [21]. There are many other telemetry stations in Russia which complement the three main antennas for near-earth missions. The Soviet Deep Space Network has successfully exchanged data flows of 120,000 bits per second (0.12 megabits) per second [21]. This network does not have global reach, and ofter cooperates with the European Space Agency.

2.4 Chinese Deep Space Network

The Chinese Deep Space Network includes multiple stations in China and five stations elsewhere in Australia, Pakistan, Chile, Namibia, and Kenya [19]. The Chinese network includes 18, 35, 40, 50, and 64 meter dishes around the world with upgrades in development at many of these sites as part of China's ambitions in lunar exploration [11].

2.5 Indian Deep Space Network

The Indian Deep Space Network is built around the Indian Space Research Organization's Telemetry, Tracking and Command Network (ISTRAC). ISTRAC is augmented by several dishes including a 32, 18, and 11 meter dishes. ISTRAC has installations on India, Russia, South America, Hawaii, and multiple islands in the Pacific and is augmented by the NASA Deep Space Network [15].

2.6 Usuda Deep Space Network

The Usada Deep Space Center has a core facility in Usada, Japan. The facility's main receiver is a 64-meter beam waveguide antenna. Japan's beam waveguide technology was adopted by NASA into their Deep Space Network. The network includes four additional locations in Japan and stations in Sweden, the Atlantic near Morocco, Western Australia, and Santiago, Chile [22].

3 DATA FLOW FROM TELEMETRY AND DEEP SPACE DATA

Each of the major space programs operates a network of ground stations which gather data. NASA's Deep Space Network operates the highest volume of data flow among these programs [2]. Unmanned spacecraft return data measured in megabits per second, which is a challenging flow of data, but not every mission poses the challenges characteristic of big data. The transmission of data becomes increasingly expensive as a space craft travels farther from earth. The Voyager 1 transmits data in real-time at a rate of 160 bits (20 bytes) per second and its 1970s era computing technology makes the data it returns expensive, and scientifically valuable, but not very big data [10]. Newer deep space spacecraft have better broadcast capabilities and can return more data when closer to the earth. The most data-productive unmanned spacecraft return data at a rate of megabits per second.

Data from these craft can grow large over the duration of a mission. For example, the Mars Reconnaissance Orbiter (MRO) [18], as of 2017, has returned over 300 terabits [9] (37.5 terabytes) of imaging data of the Martian surface. The volume of data returned by this mission is still less than what the MRO's sensors are capable of observing. The MRO carries multiple cameras with different resolution qualities. The Context Camera [9] has photographed 99.1 percent of the Martian surface and 60.4 percent has been photographed twice. The MRO also carries the High Resolution Imaging Science Experiment (HiRISE) [18] which can zoom into changes to the surface which are spotted by the Context Camera. The higher resolution of the HiRISE has limited this camera's coverage to only three percent of the surface. Additional MRO cameras photograph the entire planet each day tracking weather changes and atmospheric conditions. Although the MRO has returned almost 40 terabytes of data, the imaging instruments it carries generate significantly more data than is transmitted back to earth.

The big data challenges in space telemetry data come from the manned spacecraft, which have been exclusively earth orbiting missions since the return of Apollo 17. The International Space Station (ISS) is the collaborative mission which manned missions for the US, European, and Russian space programs have focused on since Apollo 17 returned in 1972. The ISS generates significant scientific, telemetry, life support system, and even livestreamed video from crew members [12]. The ISS has connectivity of 300 megabits per second and is continuously operating to return data [23]. Upcoming plans to add storage for additional experiments will increase the data generates and broadcast back and NASA is working to increase transmission capacity to the ISS [23].

4 DATA PROCESSING IN TELEMETRY AND DEEP SPACE DATA

Modern spacecraft exchange data via radio wave transmissions. This data is of course transmitted in binary, but receivers can be either digital or analog. Data is encoded using pulse code modulation and transferred in a way that creates a level of redundancy in the signal to maximize the quality of data received by communications networks [2]. The structure of the expected data flow from any given sensor is known before the spacecraft is ever launched. The main risk to the quality of data is noise introduced to the data stream by earth's atmosphere, temperature variations through space which the transfer traverses, or losses of signal connection during transmission. Once data is received, it is a matter of decoding from binary to the format required for any particular data stream [13, 24].

The variety of data received from spacecraft is dictated by the variety of cameras and sensors any given vessel carries. The expected data streams from any given instrument are well known and well-tested by researchers long before a spacecraft is launched. These data streams require processing once received on earth to correct for noise and missing periods of data [17]. The analysis of scientific research data is unique to each instrument and this raw data is rarely made public. The structure of this data, such as categories of data and the expected length of a dataset, is known and planned before the data is generated. With the largest datasets measured in terabytes over the course of a mission, the data is certainly large, but well within the capacity of many storage services [6]. The search and query methods created for the largest big data generators, such as social media, genetics, and astronomy, are powerful enough to perform search functions on data from unmanned spacecraft.

Telemetry data does, however, pose challenges in data variety. The ESA has created a network called 'Technology Exchange' where solutions to data processing are offered. Database management is essential since telemetry data is only of use when it is spatially and temporally noted, searchable, and quickly accessible [13].

5 BIG DATA CHANGES IN TELEMETRY AND DEEP SPACE DATA

The primary limitation preventing deep space data from becoming truly big data is the limitations posed by radio wave communication [24]. The radio receivers on earth are sensitive to the edge of the solar system. The camera definition and sensor sensitivity across the electromagnetic spectrum have improved dramatically in recent decades. The solar and nuclear power cells needed for extended deep-space travel are well developed [20]. The technological limits created by the 100 gigabit per second limit to radio wave transmissions are the bottleneck to unleashing big data in telemetry and deep space data. After decades of development and

refinement of telecommunications technology, the theoretical limits of radio-wave communication have been reached [16].

Laser communication was first demonstrated in 2013 when NASA successfully transmitted an image of the Mona Lisa [17] to the Lunar Reconnaissance Orbiter [5] and back to earth again. Laser optical technology uses wave lengths which are orders of magnitude shorter and will spread out significantly less than radio waves. Laser technology is also capable of transmitting at rates of 10 times the volume, with theoretical limits of 100 times the volume, of radio transmitters. The Laser Communications Relay Demonstration (LCRD) is an upcoming mission which will test the transmission process, encoding techniques, and atmospheric noise solutions for a laser communication system with a satellite in earth orbit [8]. The LCRD is in testing through 2017 and will be launched in 2019. Laser communication would make it possible to return data from spacecraft around the solar system in definition which would match the detail used to track environmental changes around earth.This communication method not only allows higher volume and quality of data, it can also transmit with lower power requirements and lighter transmitting equipment, which in turn allows more scientific sensors on any given spacecraft which again increases the volume of data returned [8, 24].

6 CONCLUSION

Deep space and telemetry data is on the cusp of dramatic growth in the volume, variety and quality returned by deep space spacecraft. The technological shift from radio to laser transmissions will allow an order of magnitude more data volume to be transmitted from space, whether near-earth orbit or deep space [6]. Researchers will be able to improve of definition, density, quality, and variety of data which any given deep space spacecraft will return. The analytical, storage, searchability, and visualization tools developed with big data projects will become far more applicable to deep space and telemetry data as this technology is developed. The demand for experimental access to the ISS and deep space missions is increasing every year [6, 12] and big data framework will grow more relevant as data transmission capability improves and the NASA Open Data project makes more of this data publicly available.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and allowing a focus on space travel-related data in this course.

REFERENCES

[1] Rhett Allain. 2015. Whatfis So Special About Low Earth Orbit? WIRED. (Oct. 2015). https://www.wired.com/2015/09/whats-special-low-earth-orbit/
[2] Timothy T. Pham Andrew O'Dea. 2013. Telemetry Data Decoding. *Calofornia Institute of Technology* 208, JPL D-19379 (2013).
[3] ERA. 2017. tracking network operations. Estrack Operations. (Nov. 2017). http://www.esa.int/Our_Activities/Operations/Estrack/Estrack_ground_stations
[4] NASA Facts. 2005. Deep Space Network. *Jet Propulsion Laboratory: California Institute of Technology* (2005).
[5] Jeremy Hsu. 2010. NASA to Boost Speed of Deep Space Communications. Astrobiology Magazine. (Jan. 2010). https://www.space.com/7815-nasa-boost-speed-deep-space-communications.html
[6] Amber Jacobson; Ashley Hume. 2016. NASA Communications Network to Double Space Station Data Rates. NASA Space Communications. (Dec. 2016). https://www.nasa.gov/feature/goddard/2016/nasa-communications-network-to-double-space-station-data-rates
[7] William A. Imbriale. 2002. *Large Antennas of the Deep Space Network.* Issued by the Deep-Space Communications and Navigation Systems Center of Excellence.
[8] Dewayne Washington Joshua Buck. 2013. NASA Laser Communication System Sets Record with Data Transmissions to and from Moon. nasa.gov. (Oct. 2013). https://www.nasa.gov/press/2013/october/nasa-laser-communication-system-sets-record-with-data-transmissions-to-and-from/#.Wf-6W4hry70 Release 13-309.
[9] Jet Propulsion Laboratory. 2017. Prolific Mars Orbiter Completes 50,000 Orbits. *California Institute of Technology* (2017). https://www.jpl.nasa.gov/news/news.php?release=2017-094
[10] Jet Propulsion Laboratory. 2017. Voyager Mission Status. California Institute of Technology. (Nov. 2017). https://voyager.jpl.nasa.gov/mission/status/
[11] Luan. 2015. Chinese space station is "for exclusively scientific and civilian purposes": Argentine gov't. News.cn. (June 2015). http://news.xinhuanet.com/english/2015-06/30/c_134368151.htm
[12] Bradley J. Betts; Rommel Del Mundo; Sharif Elcott; Dawn McIntosh; Brian Niehaus; Richard Papasin; Robert W. Mah. 2017. A Data Management System for Internatioja;International Space Station Simulation Tools. *Smary Systems Research LAboratory NASA Ames Research Center* (2017).
[13] Metry. 2013. Metry fi?! Telemetry and big-data framework. ESA Technology Exchange. (Jan. 2013). http://www.esa-tec.eu/space-technologies/from-space/metry-telemetry-and-big-data-framework/ Reference No. TDO0094.
[14] Don P. Mitchell. 2003. Soviet Telemetry Systems. (Dec. 2003). http://mentallandscape.com/V_Telemetry.htm
[15] Indian Deep Space Network. 2012. Indian Deep Space Network (IDSN): India's First Scientific Mission to Moon. vssc.gov.in. (Oct. 2012). http://www.vssc.gov.in/VSSC_V4/index.php/ground-segment/82-chandrayaan-1/967-indian-deep-space-network-idsn
[16] D. Herr P. Bricker, J. Luecke. 1990. Integrated receiver for NASA tracking and data relay satellite system. *IEEE Spectrum* (1990). https://doi.org/10.1109/MILCOM.1990.117374
[17] NASA Mission Pages. 2013. NASA Beams Mona Lisa to Lunar Reconnaissance Orbiter at the Moon. nasa.gov. (Jan. 2013). https://www.nasa.gov/mission_pages/LRO/news/mona-lisa.html
[18] Brid-Aine Parnell. 2017. Mars Orbiter Has Circled The Planet 50,000 Times. forbes.com. (March 2017). https://www.forbes.com/sites/bridaineparnell/2017/03/30/mars-orbiter-has-circled-the-planet-50000-times/#297356ed65fd
[19] physorg. 2011. China 'has Australia space tracking station'. Phys.org. (Nov. 2011). https://phys.org/news/2011-11-china-australia-space-tracking-station.html
[20] Jake Port. 2016. Where do deep space probes get their power? cosmosmagazine. (July 2016). https://cosmosmagazine.com/technology/where-do-deep-space-probes-get-their-power-from
[21] revolvy. 2012. Soviet Deep Space Network. revolvy.com. (2012). https://www.revolvy.com/main/index.php?s=Soviet%20Deep%20Space%20Network
[22] Nagano Saku-shi. 17. Usada Deep Space Center. Japan Aerospace Exploration Agency. (Nov. 17). http://global.jaxa.jp/about/centers/udsc/index.html
[23] Sudipto Ghoshal Ann PattersonHine Richard Alena Somnath Deb, Chuck Domagala. 2001. Remote diagnosis of the International Space Station utilizing telemetry data. *SPIE Proceedings* 4389 (July 2001). https://doi.org/10.1117/12.434253
[24] Gianluigi Liva Tomaso de Cola, Enrico Paolini. 2011. Reliability Options for Data Communications in the Future Deep-Space Missions. *IEEE Proceedings* 99, 11 (Nov. 2011), 2056–2074.

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 Leaning, 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

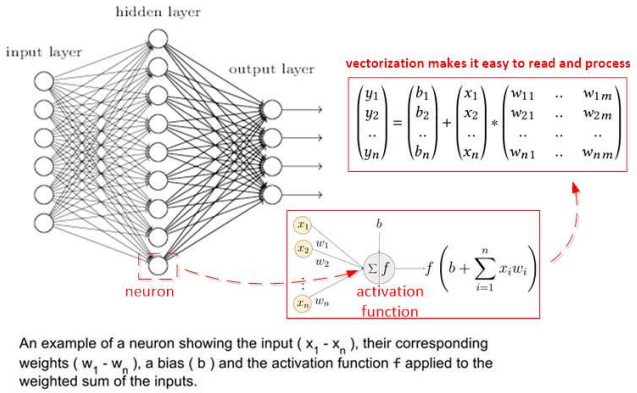


Figure 1: Simple Neural Network [3, 4]

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.

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*.

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

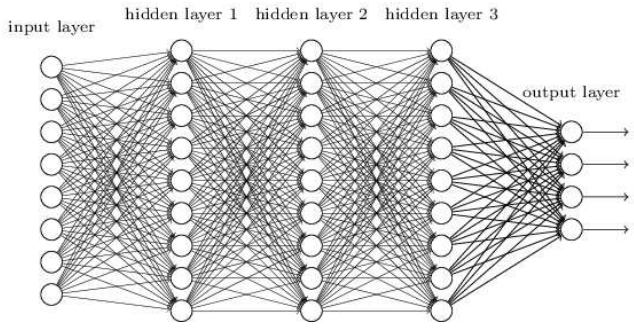


Figure 2: Deep Neural Network with three hidden layers [3]

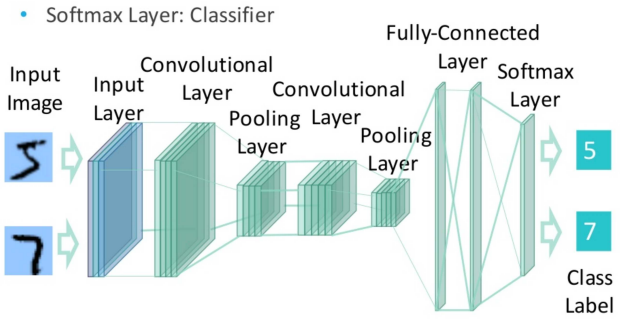


Figure 3: Sample Convolutional Neural Network [1]

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

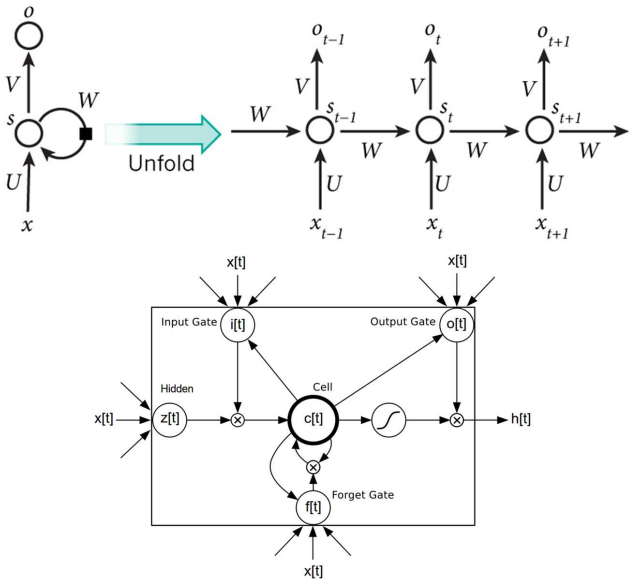


Figure 4: Recursive Neural Network Loop and LSTM Cell State [7, 8]

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.

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
[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

BigData Applications in Social Media for Marketing

Lokesh Dubey
Indiana University
3209 E 10th St
Bloomington, Indiana 47408
ldubey@indiana.edu

ABSTRACT

It has been widely accepted fact and has been addressed in many studies that Social Media data is prominently growing and diversified way of communication which entails immense possibilities with the data that it generates. It is not only a medium of socializing but a major platform for businesses to capitalize on this medium to increase their revenue in every way possible. Many new avenues of complexities and diversities introduced in social media data are identified and their respective challenges are stated. Challenges are identified in handling the social media data to derive correlations and implicit key performance indices to improve the efficacy and outreach of marketing campaigns of organizations. Big Data solutions to keep up with the paradigm shift in social media data and to remain focused towards upcoming possibilities are stated with some illustrations on political marketing and Internet Of Things.

KEYWORDS

i523, hid309, Apache Spark, Hadoop, Social Media Analytics, Marketing

1 INTRODUCTION

Social Media has become an integral part of any business working on larger scale and it has to be addressed without fail. With the advancements in technology and internet as a means of communication has drastically changed the way people interact today [15]. Merely, the scale of internet and its availability has given a lot of business owners an opportunity to not rely on meetings in person for carrying out the business. The advancements, popularity of Social Media and equally thriving industry to provide much more advanced devices to have access to internet and social media has made it a major means of communication. Most prominently social media has become a means of entertainment as well [14]. Social Media is a platform where anyone at anytime can engage with a plethora of information. The type of information shared is just about everything. It is not confined by the limitations of television or film industry where a certain body, media broadcast companies, drive what content would be shared with the audience. Here any individual can just share any kind of information which has become a phenomenon where rest of the world can participate and exponential share that information [14]. And, just like any entertainment platform when there is an engagement of a very large amount of people it becomes a very viable and extremely profitable means for marketing and advertising.

Marketing on social media is, if not completely, a very different paradigm than any other kind of marketing. There is a multitude of avenues and areas to explore in social media that are to be explored in great detail to find the right strategy. It's not that its very difficult to advertise anything on social media, in fact, social media provides so much diversity in the ways to communicate with the customers that if used intelligently organizations can benefit a great deal from it and should be able to run their marketing campaigns much more efficiently than wasting a lot of resources on areas which are wasteful [18]. The reason behind this diversity is because the sheer amount detailed information that is available on social media. Unlike other platforms which have a high social presence or are largely a vital target area for marketing the information about the consumer remains either restricted or none. From other platforms it is mostly a prediction to identify the demographics of the individual engaging on that platform. Firstly, its not just the demographics which are useful to identify an organization's correct audience [20]. It has to be much more than that, the psychographics and general attitude of the consumers is equally important to identify and minimize the subset of audience to which the advertising should be directed. Various marketing campaigns can benefit a lot from this information because it reduces the cost of marketing while increasing the rates of conversion because of an intelligent choice of audience. The question arises here is how to get this information. And social media is all about data. In every minute on social media there are 500 hours of video uploaded on Youtube, 3.3 million Facebook posts are made, 3.8 Google Searches are done, around 400,000 tweets are posted on Twitter [3]. This is what makes this a very interesting scenario, on one hand one gets a substantial amount of data daily which is a challenge to handle with any contemporary technology and on the other hand all of this data has nearly all the information that a marketing campaign would need to accurately target their consumer base. And this is where Big Data applications and contemporary technologies provide an equally interesting means of solving this problem.

2 NEW AVENUES IN SOCIAL MEDIA DATA

As we established before, social media data offers immense possibilities of how individuals can handle the data. On various studies it has been already identified and is accepted by many organizations that the data to be tackled here is huge [3]. However, just having a large amount of data is not the problem. It depends on how creatively the data can be utilized. Like other marketing mediums this is not all reliant only merely the number of views. It is much more than that and it entails practically endless amount of possibilities of what kind of data can be mined. This data can be useful on finding the demographics, psychographics, societal connections etc. In addition to that, this data can give marketing campaigns insights on things which were never even though useful before [20]. Organizations can try and delve into finding the whole lifecycle of the product being advertised right in front of them and can recursively keep improving on them. When a product is advertised for the first

time one can get feedback and sentiments about the product right away from data itself and can tune, align their marketing strategy accordingly. On the other hand it is also a challenge that this is all happening publicly [12]. And as much as this is helpful unlike any other one sided platforms where the end consumers of the product cannot engage on the platform, here the consumer can speak up and publicly announce the fallacies or even good reviews about the product.

In addition to these explicit data mining, there are many other levels on which data is can be mined for improving the performance of marketing campaigns [16] [8]. Not directly pertaining to the product at hand but all this data available on the consumers and their behaviors, social circle, likes, dislikes, can help with many other leads. They may provide a viable costumer, it can also be used to decide if the data from a certain individual is an outlier or indeed a reliable data. By looking at a behavior of a certain individual on social media and his/her interactions, information shared, social circle, likes and dislikes one can figure out of that individual is mischievous and highly unreliable or may be is simply a bot [9].

In addition to these avenues, and social media being biggest, richest and dynamic content generator, all of this we discussed is never constant. It keeps changing with new social media organizations trying to bring in more an more and features everyday and trying their best to engage more people on their sites. Which, in turn, increases the possibility of new type of data getting generated very often. One of the major paradigm shift in this direction has been Internet of Things (IOT) [2] [19]. IOT is most inadvertently changing the completely dynamics of ever changing social media data. Organizations and other businesses are trying to capitalize their level best on this new technology which helps them introduce their products like washing machines, refrigerators, vending machines on the internet. As it provides excitement to their customers as this adds more level of comfort and trying new things has always been entertaining. For instance one organization tried to open up a vending machine based on a twitter hashtag [19]. It's fun but this is an indication of on what level and how diverse this data can get in near future with IOT.

3 CHALLENGES OF SOCIAL MEDIA DATA AND MARKETING

Social media being biggest, richer and extremely dynamic makes it very big challenge for traditional technologies to tackle. However, merely the size of the data is not something new in it has been widely established the sheer amount of data in Social Media is huge and advancements, new technologies are an immediate requirement. Not only that, but these technologies needs to keep updating, refreshing themselves to keep up with the pace of new data that is being generated everyday and most importantly the type of data that keeps changing.

The most common challenges are more or less already acknowledged and there are solutions available for them or if not at least there are methodologies to handle those. Natural language processing, opinions, sentiment analysis are some common challenges which are handled with machine learning, many libraries which provide a certain level of accuracy in sentiment analysis. As far as the amount of data is concerned Big Data technologies like Hadoop MapReduce, Hive etc. are already being utilized to handle these scenarios. However, as stated before the social media space is extremely dynamic and organic space where the data is extremely volatile as far the type and life cycle of the data is concerned. Besides, social media data in itself doesn't follow any pattern. Data and the information available can change substantially based on the type of social networking platform is under the scanner. For instance Instagram[1] has very little to do with any textual information as its mostly about people posting pictures on it. There are certain ways to help image processing with the data available in form of comments and hashtags on the posts but its very limited. On the other hand sites like Facebook can be everything including, photos, textual data, audio files etc. The challenges are not only pertaining to the amount of data but how much any marketing campaigns wishes to gain from that data [11]. As there are immense possibilities of finding new insights and complex correlations which can in one way or other provided different key performance indices for increasing the efficacy of marketing campaign. And to help marketing campaigns and organizations the technical advancements and solutions are to be thought in greater detail to benefit marketing domain [4].

4 BIG DATA APPLICATIONS AND CAPABILITIES

The magnitude of social media data is not a news and there are many well defined architectures available and to an extent implemented as well which try to solve this problem in a most generic way [17]. We can agree on this partially that the amount of data and how that will be handled has been identified and there are technologies available which are fairly scalable to not pause any limitation in near future. However the bigger challenges lately have not been just the amount of data but the type of data and the methodologies which can be utilized, while not compromising on amount of data that can be handled, to get the most important and vital piece of information from the data. As stated before data in recent past is not merely about counts and views [8]. Data contains a lot more of insights and with ever changing social media market and its acceptance it is almost always volatile.

With recently growth of technologies like Apache Spark[2] and its extremely diverse set of APIs it provides a great customizable platform which can be hosted on any of the Big Data solutions like Apache HDFS [3] for distributed computing to ensure that organizations are getting the best out of there data and are not lagging behind because of technological limitations. There are more details provided later for some sample solutions where Spark can be used but for now we can draw some parallels on what kind of useful performance indices can be retrieved from social media data and how Big Data solutions can be useful there. Apart from the size of the data and the reception of a particular post or an ad in any social site there are lot of works done around natural language processing for sentiment analysis of the data [13]. However, these details mostly pertain only to opinion mining, sentiment analysis of

[1] Instagram is a mobile, desktop, and Internet-based photo-sharing application and service that allows users to share pictures and videos either publicly or privately
[2] Apache Spark is an open-source cluster-computing framework
[3] Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model.

how the posts or products were perceived by the audiences [6]. But there's more to this data. Various libraries of Spark like Machine Learning Library[4] and GraphX[5] can be utilized to build models which are not merely dependent on sentiments or opinions but to mine data on how the product is being perceived by the consumers as a group or as a societal importance. Which can of course be related to the preexisting sentiment analysis by natural language processing but this provides an acute subset to work with as this can minimize the target audience to a certain extent.

Twitter[6] tweets can be collected and related to each other with a GraphX library and models can be created to find out what causes anything to go on a trending page or what are the nitty gritties that can get consumers of twitter more excited about any product which of course increases the footprint of any advertisement more. This helps however only with increasing the outreach of the particular advertisement. Or on the other hand it can help create a model which can identify, by the interactions of different user with other users, to identify social media bots which are not a potential customer for the organizations which should be weeded out while considering and calculating any kind of the key performance indices by a marketing campaign.

5 POLITICAL MARKETING

As much as marketing of commercial products has its own space in social media, it has a very huge impact on political marketing as well. And there are two sides to this which were exacerbated and surfaced in recent US presidential elections. Marketing on social media for any political campaign has two challenges. Primarily it is to promote one particular individual like any other traditional marketing but on the flip side of it one has to make sure that there are no entities are utilizing the same platform to alter the results of the election by utilizing the openness of social media and promoting fake information which most of consumers can readily believe in without ensuring the authenticity of the posts. Which in traditional terms was also considered as echo chambers or filter bubbles where they are kept aloof from contrary perspectives and possibly reality. Post 2016 US presidential elections it was researched that a lot of fake Facebook posts were circulated from within the country and even from other countries [1] [10]. This represents a new and a unique challenge for the marketing world as this is totally different type of data to deal with, compared to commercial marketing and also its a unique type of data mining.

In this particular scenario even for a giant organization like Facebook took around 6 to 7 months to identify the authenticity of the claim of fake news and to weed out and identify the exact ads and posts which were not authentic [7]. However, this can be done proactively by the political campaigns proactively to ensure that no unethical activity on social media can affect their campaigns.

Data science methodologies and a spark platform can provide a lot of insights on this particular field. Most of this data is generated and fake accelerated in form of making it looking like a trending topic with endorsements showing that many people like it and that it was shared by numerous people and it's done by social media bots [21]. Many of the bots can be utilized to create fake debates on fake posts which can also be perceived by the audience as firstly true and secondly being endorsed by a lot of other social media users which can help tilting or at least affecting the decision of a voter [22].

To avoid this with all the data available on social media is crucial and helpful. All the contemporary key indices which are retrieved from the data today in form of sentiment analysis, opinion mining, counts of reposts, share can play a very vital role in this kind of data mining but to a larger extent this information also needs to be supported by behavioral information which can be retrieved from the data by building correlation between the various posts and data that is shared by the social media users. Spark and its library can be used to build graph based models to generate and weed out the social media handles/accounts which seem to be posting unethical information and will help to weed out those kinds of accounts to avoid swaying away any voter's decision in wrong direction.

6 INTERNET OF THINGS

Imagine one individual's washing machine posting a Tweet on manufacturer's page publicly about a certain anomaly. Or may be posting stats of its last week run and how much water it saved or how much electricity it wasted. Technology world is on the verge of two possibly biggest data generators merging together. Devices at a consumers home may remain constantly connected with its owners even while he's out of town. They might be able to connect to other devices in the vicinity and possibly identify if the issue it is facing to function is isolated or a general problem in the community. Essentially, we are looking at a dawn of endless possibilities as initially was thought about social media.

There are many aspects to IOT which can exponential increase the complexity and richness of the data available on social media. On one hand this may also, as it's usually the case with any technology, pause a threat on privacy etc. but this will continue to thrive on the basis of providing at most comfort level to the consumer of devices [5]. The variations in the richness of data and complexity would be endless. There could be patients and their devices altering their doctors directly on their social media handles. The information can be broadcasted on multiple groups probably for the hospital employees to ensure their patients are constantly monitored even when its not physically possible. Every now and then certain group of devices like refrigerators or microwave ovens of a certain brand may start showing up on a trending page because they have been saving so much enery lately and they have received rave reviews about them on social media by their owners.

Yet again, we're talking about the endless amount of rich and diverse data which opens up the doors for numerous of ways in which this data can be utilized. With IOT organizations can utilize the data available from the devices and their posts on social media to garner statistics on how well their devices are working, how good is a reception of the devices, in fact, this might even be recursive. In every which way, the quest for enriching lives of humans would continue and by learning from the data posted by the devices and their consumers more avenues of possibilities will keep opening to make it better and better each day.

[4]MLlib is Apache Spark's scalable machine learning library.
[5]GraphX is Apache Spark's API for graphs and graph-parallel computation.
[6]Twitter is an online news and social networking service where users post and interact with messages called tweets

Big Data technologies very soon will be utilized to make this possible. Today, IOT is mostly focused towards data that is posted directly from the devices to certain servers of the manufacturer or the maintenance and it is indeed large enough to be requiring Big Data technologies to handle it. However, with social media this will give another good use case to make predictions. Let's say based on the social impressions of the owners of washing machines complaining a lot about their machine's not working in the optimal way lately to their friends and many washing machines of that particular area possibly posting an collaborating on certain groups on social sites about water being to hard lately may help identify the issues and identify in exact which locality this is happening and will help rectify the issue permanently as soon as it starts to appear. This exercise can again be fed back the system and marketed as level of comfort and proactive rectification their devices provide to the customer in form of more social media impressions which may in turn also increase there consumer base.

7 CONCLUSION

Various different proliferating avenues in social media that are changing the diversity, richness and complexity of the data available on social media that can help marketing campaigns and various organizations in many different ways. This data is so enriched and perplexing that marketing industry can choose based on their criteria on how much they wish to delve into this data. More information is never bad but handling that data and being able to mine it for different key performance indices can be made possible with the contemporary Big Data technologies. Growing and organically changing social media data might have made it difficult for various traditional approaches of identifying the right strategies of marketing but with latest Big Data technologies like Spark Streaming, MLib, GraphX it is possible to tackle any kind of data in the most fluid form so that it is customizable enough to quickly adapt based on the changing needs of social media data. There are many challenges on handling this new diversity and complexity of this data in form of more detailed information that is available other than simple number of views or demographics. It is established that the availability of data and insights to find complex correlations between the data is practically endless and it relies highly on the will and to the extent marketing campaigns wish to increase their efficacy of their marketing which makes this more complex and requires more creativity. Almost all the type of data that is being generated and is available has been well thought of and aligns well with technology available in Big Data. However, it has been done mostly in social media and Big Data spaces individually. There's a high need to focus on more solutions which are more targeted and focus towards the major paradigm shift and viral behavior of social media. Which is also at the brink of increasing and diversifying immensely after various different new usages of social media like its use in politics and Internet Of Things.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election.* Working Paper 23089. National Bureau of Economic Research. https://doi.org/10.3386/w23089

[2] Robert Allen. 2016. 7 examples of marketing applications of the Internet of Things which are here now. (2016). https://www.smartinsights.com/managing-digital-marketing/marketing-innovation/7-examples-applications-internet-things-now/ accessed 2017.

[3] Robert Allen. 2017. What happends online in 60 seconds? (2017). https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/ accessed 2017.

[4] Alexandra Amado, Paulo Cortez, Paulo Rita, and Srgio Moro. 2017. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics* 34 (2017), 1–7. https://doi.org/10.1016/j.iedeen.2017.06.002

[5] Lawrence Ampofo. 2014. 5 Ways The Internet Of Things Will Change Social Media. (2014). https://www.business2community.com/social-media/5-ways-internet-things-will-change-social-media-01047822#VXXKbHBWmiOEyxjT.97 accessed 2017.

[6] Bogdan Batrinca and Philip C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY* 30 (01 Feb 2015), 89–116. https://doi.org/10.1007/s00146-014-0549-4

[7] Cyrus Farivar. 2017. Facebook apologizes for allowing Russian ads to interfere with 2016 campaign. (2017). https://arstechnica.com/tech-policy/2017/10/facebook-execs-are-talking-about-russian-ads-but-theyre-not-saying-much/ accessed 2017.

[8] Mylynn Felt. 2016. Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society* 3 (2016), 2053951716645828. https://doi.org/10.1177/2053951716645828

[9] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59 (June 2016), 96–104. https://doi.org/10.1145/2818717

[10] David Folkenflik. 2017. Facebook Scrutinized Over Its Role in 2016's Presidential Election. (2017). https://www.npr.org/2017/09/26/553661942/facebook-scrutinized-over-its-role-in-2016s-presidential-election accessed 2017.

[11] C.F. Hofacker and D. Belanche. 2016. Eight social media challenges for marketing managers. *Spanish Journal of Marketing - ESIC* 20 (2016), 73 – 80. https://doi.org/10.1016/j.sjme.2016.07.003

[12] Ines Schulze Horn, Torben Taros, Sven Dirkes, Lucas Hüer, Maximilian Rose, Raphael Tietmeyer, and Efthymios Constantinides. 2015. Business reputation and social media: A primer on threats and responses. *Journal of Direct, Data and Digital Marketing Practice* 16 (01 Jan 2015), 193–208. https://doi.org/10.1057/dddmp.2015.1

[13] Merin K Kuriani, Vishnupriya S, Ramesh R, Divya G, and Divya D. 2015. BIG DATA SENTIMENT ANALYSIS USING HADOOP. *International Journal for Innovative Research in Science & Technology* 1, 11 (2015), 92–96. http://www.ijirst.org/articles/IJIRSTV1I11036.pdf

[14] Chei Sian Lee and Long Ma. 2012. News Sharing in Social Media: The Effect of Gratifications and Prior Experience. *Comput. Hum. Behav.* 28, 2 (March 2012), 331–339. https://doi.org/10.1016/j.chb.2011.10.002

[15] Hajli MN. 2014. Social media analytics: a survey of techniques, tools and platforms. *International Journal of Market Research* 56 (2014), 387–404. https://www.mrs.org.uk/ijmr_article/article/101805

[16] Michele Nemschoff. 2014. Why Marketers Love Big Data & Hadoop. (2014). https://www.socialmediatoday.com/content/why-marketers-love-big-data-hadoop accessed 2017.

[17] Ekaterina Olshannikova, Thomas Olsson, Jukka Huhtamäki, and Hannu Kärkkäinen. 2017. Conceptualizing Big Social Data. *Journal of Big Data* 4 (25 Jan 2017), 3. https://doi.org/10.1186/s40537-017-0063-x

[18] Holly Paquette. 2013. Social Media as a Marketing Tool: A Literature Review. (2013). http://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1001&context=tmd_major_papers accessed 2017.

[19] Neil Patel. 2015. How The Internet Of Things Is Changing Online Marketing. (2015). https://www.forbes.com/sites/neilpatel/2015/12/10/how-the-internet-of-things-is-changing-online-marketing/#b2680b868803 accessed 2017.

[20] Alexendra Samuel. 2016. Psychographics Are Just as Important for Marketers as Demographics. (2016). https://hbr.org/2016/03/psychographics-are-just-as-important-for-marketers-as-demographics accessed 2017.

[21] L. Weng, F. Menczer, and Y.-Y. Ahn. 2013. Virality Prediction and Community Structure in Social Networks. *Sci. Rep.* 3 (2013), 1–30. https://doi.org/10.1038/srep02522

[22] Andrej Zwitter. 2014. Big Data ethics. *Big Data & Society* 1, 2 (2014), 2053951714559253. https://doi.org/10.1177/2053951714559253

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 or 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].

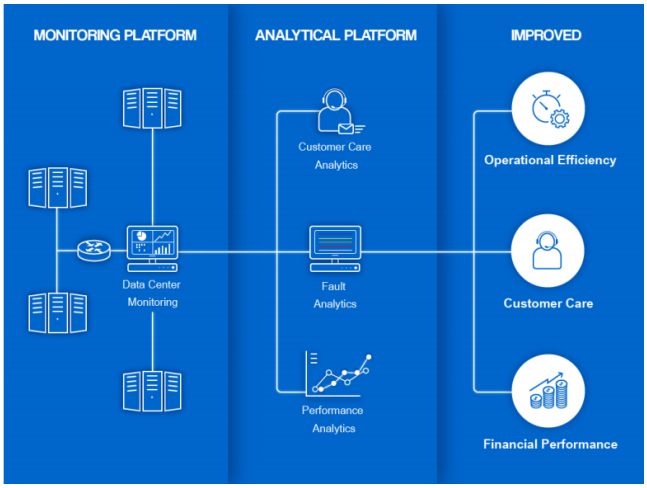


Figure 1: Monitoring and analytical platform [3]

As shown in Figure 1, data center monitoring platform is capable of monitoring all types of network devices. The events captued 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 (Service Level Agreement) 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,vitualization, 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]. The basic areas getting monitored on servers and network devices as part of fault monitoring are, server and network alarms, events, event enrichment and automatic notifications.

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.

The basic areas getting monitored as part of performance monitoring are, network performance, server and network performance threshold and 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 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

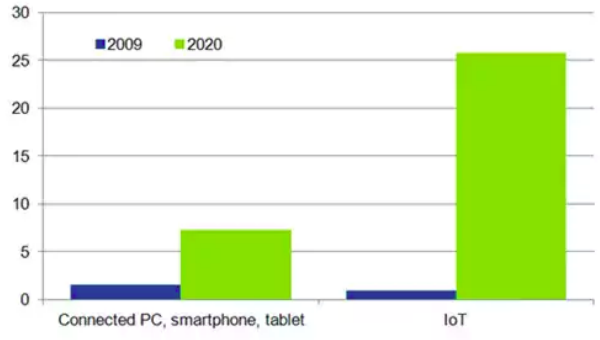


Figure 2: Total of connected devices (in billions) [13]

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, re-transmits, out of order packets, and packet fragments based on actual application traffic flows, offering the most relevant and actionable intelligence [1].

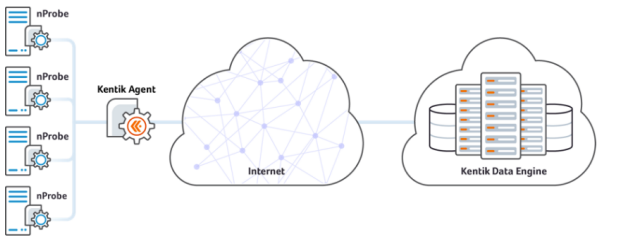


Figure 3: Big data based SaaS NPM solution [10]

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 may execute on performance data to, examine performance patterns, find commonalities, derive actionable insights. The actionable insights that can be derived are, identifying devices exhibiting deviation from normal pattern, predict network performance and load balancing requirements.

3.2 Server or 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].

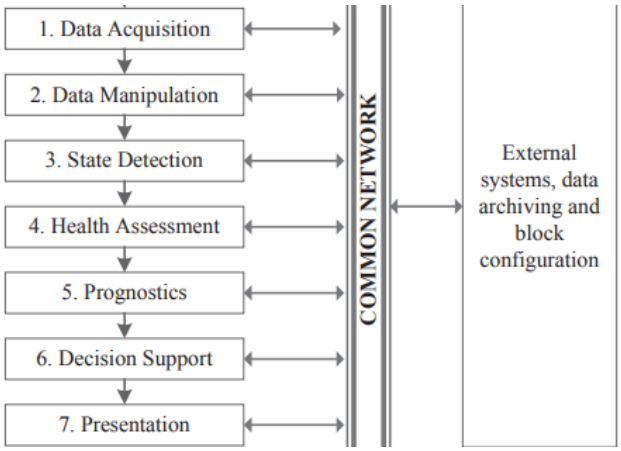


Figure 4: Fault detection architecture [14]

Figure 4, shows the network fault analysis architecture using big data. The procedure include activities like, 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 big data classification algorithm, we can predict the customer satisfaction index and the possibility of a customer to churn out of business. Clustering algorithms are effective in predicting new customer plans and their network and 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.

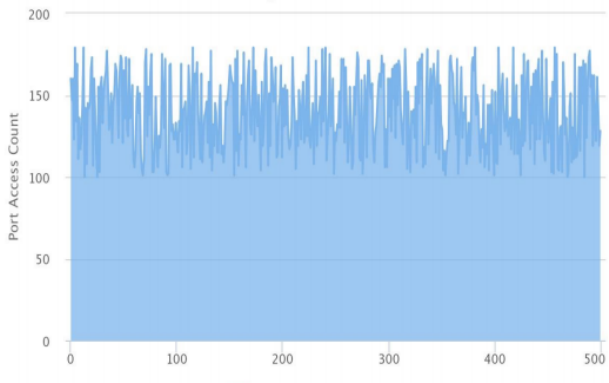


Figure 5: Ports access count within range 100 to 180 per minute [2]

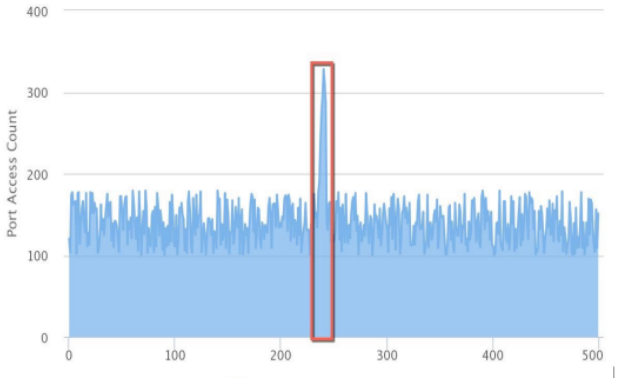


Figure 6: Ports access count shows abnormal port activity [2]

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 networks pose 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.ingrammicroadvisor.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.ingrammicroadvisor.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

MQTT for Big Data and Edge Computing

Janaki Mudvari Khatiwada
Indiana University, Bloomington
P.O. Box 1212
Bloomington, Indiana 43017-6221
jmudvari@iu.edu

ABSTRACT

With increasing use of sensors and smart devices or internet of things that generate real time data and opt for immediate output, immediate message delivery and reliable result is a must. This requires reliable efficient connection among sensors and near end devices. MQTT protocol establishes the connection among participating clients. Traditionally connected devices and industrial automation equipment are connected to cloud. Data flow between cloud services and internet of things may not be reliable and efficient all the time. With growing number of smart devices, internet of things and their applications, communications between them and cloud services need reliable internet connection with no data limitations. Edge computing, data flow between near end devices, under MQTT can be a solution to improve latency of data flow, better scalability and taking load off on local servers.

KEYWORDS

i523, hid330, MQTT, Big Data, Edge Computing

1 INTRODUCTION

MQTT, MQ telemetry transport protocol is an open source, a light–weight reliable publish and subscribe messaging protocol on top op TCP/IP protocol. It's use is intended for wireless and low band–width connection. It acts as a broker between clients or a server. Both publisher and subscriber are MQTT clients. It was invented in 1999 by Andy Stanford–Clark and Arlen Nipper. It was designed for communicating messages between clients in remote locations that had limited network bandwidth. We have come a long way since then with all new sensors and smart devices and internet of things (IoTs) that are live streaming data. These devices are connected to the cloud through sensors. Flow of data from cloud to these devices or vice–versa may be latent because of slower connectivity and depend upon the devices' performance. However, these internet of things require efficient flow of information in an instant. Reliable flow of these data among IoTs need reliable connection between server and its clients which may need higher band–with server. Servers should be able to convey messages subscribed by their clients within milliseconds. This is when edge computing comes into play. Edge computing is data analysis nearby these devices so that there is less load on the main server. All clients or smart devices are connected to the broker and broker publish messages to each client based upon their subscription topic. Popular MQTT broker such as Mosquito, HiveMQ also support web sockets so that a web page of a modern web browser can connect to the MQTT broker and can subscribe on a topic and display real time data. MQTT supports transport layer security and authenticate with username and password. Raspberry Pi, esp8266 are small devices or small computer that have built–in support for wireless LAN which can be employed as an edge device, after being integrated with MQTT, for data transmission

Why are we discussing Edge computing and MQTT? While MQTT was in use since 1999, it gained its significance recently. As with the development of IoTs and their newer applications, MQTT is seen as a light weight open source protocol that can establish a connection between server and its near end devices and help in quick message delivery. The number of connected devices are expected to reach 50 billion by 2020, from 500 million in 2003 to 12.5 billion in 2010. These sensors and devices generate huge volume of data and send to the cloud in a great velocity. This big data transmission to the cloud, which is a centralized data center, puts all the computation load on cloud and result in latent message transformation. Most of the applications in these smart devices need real time analysis and immediate message delivery on the topic they subscribe. Edge computing basically allows decentralization of data flow between these devices, servers and users. Further, edge computing significantly takes load off of cloud computing. Edge computing architecture facilitates data transmission between sensors and connected devices to a nearby edge computing device, "like a gateway device that processes or analyzes the data locally, rather than sending it to the cloud or a remote data center" [3]. A gateway could be a server or a feature of the router or a computer that helps connect devices with the data–center network. Figure1 below gives a general idea of message transmission through MQTT protocolo. In this figure HiveMQ is an open source MQTT broker.

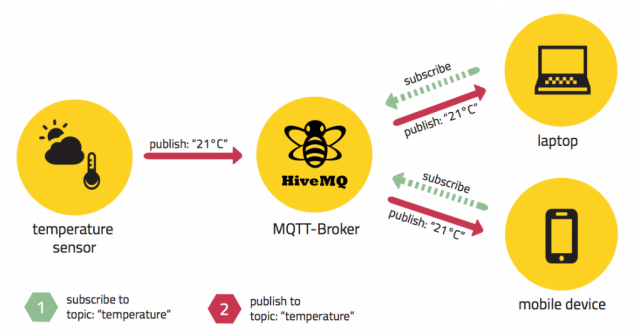


Figure 1: Pub/Sub MQTT

2 APPLICATIONS

A general mobile app such as facebook messenger that uses MQQT server for sending and receiving messages using MQQT library is

a simple example of MQQT Application. This application runs on smart devices like smartphones and tablets. Remote monitoring of manufacturing plants, smart grid, transportation, traffic monitoring, surveillance cameras,emergency response, industrial IoTs, smart home automation, Amazon Web Services, Facebook messenger are some of the notable applications of MQTT protocol for edge computing. According to an estimate by Cisco, the amount of "data generated by people, machines and things will be 600 zetta–bytes(ZB) which is up from 145 ZB generated in 2015" [3].

Industries can benefit by using MQQT protocol for edge computing since the devices and sensors may be located in remote locations. Remote locations might have network constraints such as lower connectivity, data limitations and and high latency. Given its lightweight feature MQTT protocol is ideal for data transmissions under such circumstances. Edge computing provides platform to the industries, social and government organizations so that they are able to analyze required data at the right location in the network in real time. By adopting their operations to the platform of edge computing industries could make use of these data not only to increase their profits but also enhance their operational efficiency. "Ignition Industrial Internet of Things provides an MQQT architecture platform to industries and business applications" [1].

Monitoring patient at home is a good use case of application of MQTT protocol. Patient with implanted cardioverter defibrillator, which communicates with the hospital, can be monitored remotely along with the implanted device. Device's and patient's data are transmitted to the MQTT device located in patient's residence. Data is transmitted to the hospital by a transmitter that helps to analyze the condition of the patient. This helps to reduce hospital visits [2]. It also detects emergency and notify on–call physician. The whole monitoring system is simple on patient's part as the device is already configured for MQQT client implementation by the the supplier or developer. Patient just needs to plug it in for power supply.

Another use case of MQQT protocol for big data and edge computing is environment monitoring for which sensors are most likely located in remote locations with intermittent network connection. Environmental sensors depending on the topic such as water level sensor, need to constantly stream data from the location and others such as pollutant sensors store data in in their system and transfer them in a timely manner. Home automation is another very good example of above discussed topic. Single broker or MQTT server establishes connection with smart devices around the house such as coffee–maker, air conditioner, lights, doors and others whenever required they are able to publish messages to the client upon subscription to the topic. The figure 2 mentioned below sums of the MQTT for big data and edge computing.

3 MQTT WORK–FLOW

The work flow of MQTT starts by establishing connection of a broker, which is a local server, to a client. Clients are devices connected to a broker. There is at least one MQTT client responsible for publishing a message and second client which receives the message has to be subscribed to the topic in which first client is publishing a message. A message can be a data or command. Topic is a string separated by slashes. For example, "home/kitchen/light1"

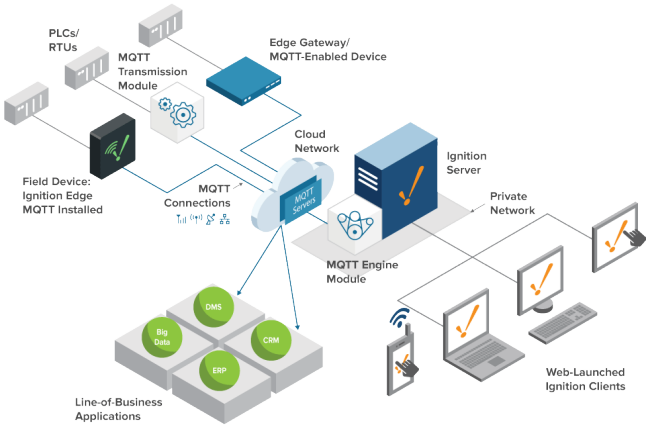


Figure 2: MQTT Architecture for Edge Computing

is a topic. Then client can also subscribe to a topic to a different device through a different command. MQTT client and server communicate through different control packets as shown in figure 3.

Control packet	Direction of flow	Description
CONNECT	Client to Server	Client request to connect to Server
CONNACK	Server to Client	Connect acknowledgment
PUBLISH	Client to Server or Server to Client	Publish message
PUBACK	Client to Server or Server to Client	Publish acknowledgment
PUBREC	Client to Server or Server to Client	Publish received (assured delivery part 1)
PUBREL	Client to Server or Server to Client	Publish release (assured delivery part 2)
PUBCOMP	Client to Server or Server to Client	Publish complete (assured delivery part 3)
SUBSCRIBE	Client to Server	Client subscribe request
SUBACK	Server to Client	Subscribe acknowledgment
UNSUBSCRIBE	Client to Server	Unsubscribe request
UNSUBACK	Server to Client	Unsubscribe acknowledgment
PINGREQ	Client to Server	PING request
PINGRESP	Server to Client	PING response
DISCONNECT	Client to Server	Client is disconnecting

Figure 3: MQTT publish Subscribe Architecture

MQTT broker receives all the messages, filters them and publishes them to each subscribed clients. one of the important feature of this protocol is it decouples devices and applications before transmitting messages or data. Each MQTT message contains a topic, payload which may be in any format like binary data or plain text

or xml. Quality of service is another feature of MQTT. It has three standard quality of services (QoS) for message delivery; 0, 1, 2. QoS 0 means message is delivered at most once, QoS 1 means message is delivered at least once, so duplicate message is possible and QoS 2 means message is delivered exactly once and there will be no duplicates but message is guaranteed.

For securing the data transmission both MQTT broker and clients authenticate each other by providing username and password during transmission and application level. Both broker and clients transmit acknowledgement message after connection, subscription and publishing of message. Another feature of MQTT is 'Last Will and Testament', which is used to notify other clients when a client is suddenly disconnected. Each client can specify its last will as a message in certain topic with quality of service, message retained and payload features. Message specified in last will and testament will be send to all the clients upon disconnection of a client [4]. Below is a demonstration of MQTT publish and subscribe python codes:

Program 4 shows MQTT host establishing connection with the client then publishing a message on topic house/light1 with Qos 0 and and QoS 1.

4 FUTURE APPLICATIONS

With increasing number of inter connected sensors and applications of smart devices and their demand by general public, there is no doubt MQTT, big data and edge computing is the next big phenomena. Similarly, industrial internet of things and their use is rapidly growing. MQTT being a light weight open source protocol with all of its features discussed as above is being used as a broker for industrial Iots. Developed countries have vision for smart cities, self driving cars, smart energy management, smart transportation and so on and so forth.

5 LIMITATIONS

MQTT topics are structured in a hierarchy similar to folders and files in a file system using forward slash as a delimiter. Topic names are case sensitive and must consists of at least of one character to be valid. Topics are not permanent and are created by publishing and subscribing client. A client can publish to only one topic. To publish a message to two topics message has to be published twice. Configurations can go wrong very easily. Since edge computing requires machine to machine communication at the edge they are prone to get compromised by hackers.So, security is a great concern. While scalability is one of the benefits of use of MQTT broker in edge computing, it is also a challenge when millions of connections are involved. This might need a group of distributed broker nodes. Edge devices have lesser capabilities for data analysis, so there might be latency since data has to be transmitted to the cloud service.

6 CONCLUSION

There is tremendous application prospects of using MQTT protocol in big data and edge computing. Edge computing, being a new development in data analytics, has enormous use as the volume of data is increasing at such an unimaginable rate. MQTT is simple to use and configure, it is a great platform for edge computing.

```
import paho.mqtt.client #paho.mqtt is a mqtt library
 import time
 broker = 'IP address of broker'
 port = 1883 #standard MQTT port
 def on_log(client, userdata, level, buf):
     print(buf)
 def on_connect(client, userdata, flags,rc)
    if rc == 0:
       client.connected_flag=True
        print("connected OK")
     else:
        print("Bad connection Returned code=", rc)
        client.loop_stop()
 def on_disconnect(client, userdata, rc):
     print("client disconnected OK"),
 def on_publish(client, userdata, mid):"
     print('In on_pub callback mid=', mid)
 mqqt.Client.connected_flag= false"
 client = mqqt.client(' ')
 client.on_log=on_log
 client.on_connect = on_connect,
 client.on_disconnect = on_disconnect
client.connect(broker, port) # establish connection,
 client.loop_start()
 while not client.connected_flag:
     print('In wait loop')
     time.sleep(1)
 time.sleep(3)
 print('publishing')
ret=client.publish('house/light1', 'Testmessage 0', 0)
 print('published return=', ret)
 time.sleep(3)
ret=client.publish('house/light1', 'Testmessage 1',1)
 print(published returned, ret)
 time.sleep(3)
 client.loop_stop() #stop the loop
 client.disconnect() #disconnect client
```

Figure 4: Program MQTT

All the commercial, industrial, service and development sectors are moving towards modernization with regard to using newer technology. Newer technologies are being developed in some way as a medium of making business, industries health care, military and other sectors efficient. Iots newer applications demand edge computing for quicker data analysis and MQTT as a simple to use protocol.

ACKNOWLEDGMENTS

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

Control packet	Direction of flow	Description
CONNECT	Client to Server	Client request to connect to Server
CONNACK	Server to Client	Connect acknowledgment
PUBLISH	Client to Server or Server to Client	Publish message
PUBACK	Client to Server or Server to Client	Publish acknowledgment
PUBREC	Client to Server or Server to Client	Publish received (assured delivery part 1)
PUBREL	Client to Server or Server to Client	Publish release (assured delivery part 2)
PUBCOMP	Client to Server or Server to Client	Publish complete (assured delivery part 3)
SUBSCRIBE	Client to Server	Client subscribe request
SUBACK	Server to Client	Subscribe acknowledgment
UNSUBSCRIBE	Client to Server	Unsubscribe request
UNSUBACK	Server to Client	Unsubscribe acknowledgment
PINGREQ	Client to Server	PING request
PINGRESP	Server to Client	PING response
DISCONNECT	Client to Server	Client is disconnecting

Figure 5: MQTT Publish Subscribe Architecture

REFERENCES

[1] Inductive automation. Undated. The Industrial Internet of Things: It's Here and It Works It's Ignition. web page. (Undated). https://inductiveautomation.com/solutions/iiot#

[2] IBM Knowledge Center. 2017. *Telemetry use case: Home patient monitoring.* Web Page. IBM. https://www.ibm.com/support/knowledgecenter/SSFKSJ_8.0.0/com.ibm.mq.pro.doc/q002780_.htm

[3] Andrew Foster. 2017. Why the Industrial IoT Needs an Open-Source Edge Platform. Web Page. (July 2017). https://www.rtinsights.com/why-the-industrial-iot-needs-an-open-source-edge-platform/

[4] HiveMQ. 2017. MQTT Essentials Part 9: Last Will and Testament. blog. (2017). https://www.hivemq.com/blog/mqtt-essentials-part-9-last-will-and-testament

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.

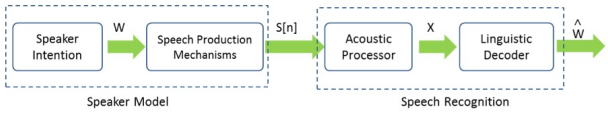


Figure 1: Conceptual Model of Speech Production

- 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

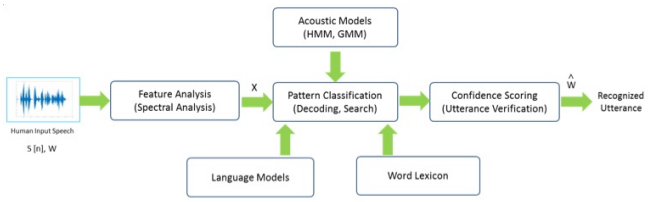


Figure 2: Speaker Recognition Concept

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 [8]. The input human speech signal, s[n], is converted to the series of feature vectors, X = [x1, x2,..., xT], 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 five 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.

Disclosure Integration: This phase involves resolving of references between sentences. Meaning of one sentence may depend on the meaning earlier sentence. Also meaning of current sentence may influence meaning of following sentences.

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 [10].

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 MaltParser, 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

Neural 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 [7] 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 [9].

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

[2] Dr E. Chandra and K. Manikandan M. S. Kalaivani. 2014. A Study on Speaker Recognition System and Pattern classification Techniques. (2014). https://www.ijireeice.com/upload/2014/february/IJIREEICE1H__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

[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] Google LLC. 2017. (2017). https://cloud.google.com/speech/

[8] 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

[9] 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

[10] KOVAR Vojtech. 2014. Automatic Syntactic Analysis for Real-World Applications [online]. (2014). https://nlp.fi.muni.cz/~xkovar3/thesis.pdf

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.

A comparative study of Kubernetes and Docker Swarm and Advantages of Singularity Container to HPC World

Anand Sriramulu
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
asriram@iu.edu

ABSTRACT

To discuss on the Container orchestration and comparing the most popular orchestrations Docker Swarm and Kubernetes on the areas of provisioning, configuration, discovery, monitoring, administration, rollback and placement policies. Providing high level overview on the advantages of the Singularity Containers over the others and how it benefits HPC workloads.

KEYWORDS

i523, hid338, Docker Swarm, Kubernetes, Singularity

1 INTRODUCTION

Technology professional see the advantage of using a microservices architecture, wherein the application comprises of loosely coupled components, such as load balancers, caching proxies, message brokers, web servers, application services, and databases. The use of microservices allows developers to quickly create applications. In addition, this architecture saves a tremendous amount of resources in scaling applications, since each component can be scaled separately. Containers make it easy to deploy and run applications using the microservices architecture. They are lighter-weight compared to VMs and make more efficient use of the underlying infrastructure. Containers are meant to make it easy to scale applications, meet fluctuating demands, and move apps seamlessly between different environments or clouds. Container orchestration tools can provide placement, scheduling, deployment, updates, health monitoring, scaling and failover functionality. [2]

2 CONTAINER ORCHESTRATION FUNCTIONS

Here are some of the capabilities that a modern container orchestration platform will typically provide:

2.1 Provisioning

Container orchestration tools deals with provisioning or scheduling containers within the cluster and launching them. This process involves determining the appropriate host for the placement of the containers based on different constraints like resource requirements, location afinity, etc. The underlying goal is to increase utilization of the available resources. Apart from a container-provisioning API, orchestration tools will invoke the infrastructure APIs specific to the host environment. [8]

2.2 Declarative Configuration

Container orchestration tools provide options for the DevOps team to define the blueprint for an application workloadand its configuration in a standard schema using languages like YAML or JSON. The definitions includes informations related to repositories, networking, storage and logs to support the workload. Defining the blueprint in this manner makes it easy for DevOps teams to edit, share and version the configurations and provide repeatable deployments across development, testing and production. [1]

2.3 Service Discovery

Container discovery becomes critical as the containers are running on multiple hosts in a distributed deployment environment. Web servers need to dynamically discover the web servers and the load balancers need to discover and register the web servers. Orchestration tools uses a distributed key-value based store as lightweight DNS mechanism to discover the containers.

2.4 Monitoring

Container orchestration tools are aware of the configuration to track and monitor the health of the containers and hosts in the cluster. If a container crashes, a new one can be spun up quickly. If a host fails, the tool can relocate the failed containers on another host. It will also run specified health checks at the appropriate frequency and update the list of available nodes based on the results. The tool will ensure that the deployment matches the desired state of the cluster matches the configuration specified.

2.5 Rolling Upgrades and Rollback

Some orchestration tools can perform 'rolling upgrades' of the application where a new version is applied incrementally across the cluster. Traffic is routed appropriately as containers go down temporarily to receive the update. A rolling update guarantees a minimum number of "ready" containers at any point, so that all old containers are not replaced if there aren't enough healthy new containers to replace them. If, however, the new version doesn't perform as expected then the orchestration tool may also be able to rollback the applied change.

2.6 Policies for Placement, Scalability etc.

Container orchestration tools provide a way to define policies for host placement, security, performance and high availability. When configured correctly, container orchestration platforms can enable organizations to deploy and operate containerized application workloads in a secure, reliable and scalable way. For example, an

application can be scaled up automatically based on CPU usage of the containers.

2.7 Administration

Container orchestration tools should provide mechanisms for administrators to deploy, configure and setup. An extensible architecture will connect to external systems such as local or cloud storage, networking systems etc. They should connect to existing IT tools for SSO, RBAC etc.

3 KUBERNETES

Kubernetes is a Goolge product and they used it for their running their heavy workloads in production. Kubernetes website says, "Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications." It an architecture based on master server and multiple nodes(minions). To manage and orchestrate the nodes, kubecfg (Command line tool) is used to connects to the API endpoint. Definitions of the components within the Kubernetes environment:
Master: The server that runs the Kubernetes processes like API service, scheduler and replication controler.
Node: The hosts that runs the kubelet service and the containers engine. The node receives command from the Master.
Kubelet: It's a node lever manager.
Pod: Collection of containers deployed to the same node.
Replication Controller: Defines the number of pods or containers need to be running.
kubecfg: Command line interface to manage the kubernetes deployment.
[3]

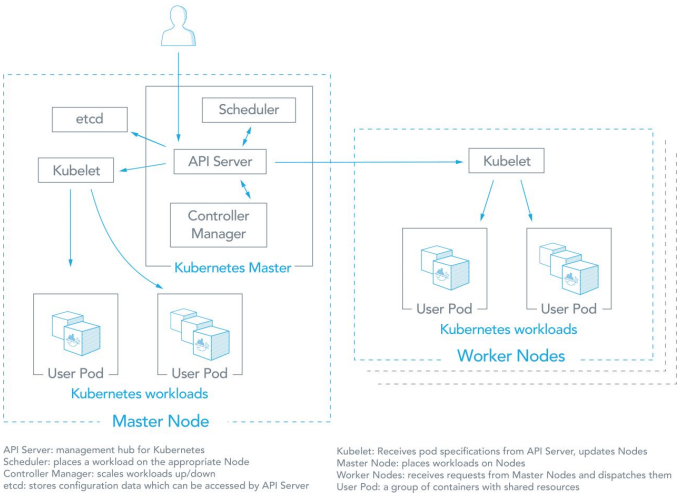


Figure 1: Kubernetes Nodes Illustration [5].

4 DOCKER SWARM

Docker swarm is a simple container orchestrion and yet it's powerful. It uses the same Docker API with the core Docker engine. The swarm manages pool of Docker engines with an endpoint in which it benefits the existing tools and APIs as it work with the cluster the same way it work with a docker instance. The Docker Swarm has in-built scheduling strategies as the containers can be placed across the cluster and also it supports random placement. Swarm uses a pluggable back-end architecture that works with a simple hosted discovery service, static files, etcd, Zookeeper, Consul,lists of IPs.

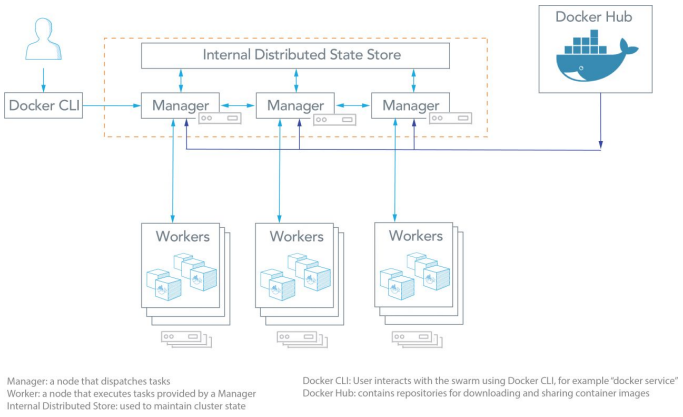


Figure 2: Docker Swarm Architecture [5].

5 KUBERNETES VS DOCKER SWARM

Both the contrainer orchestration offers similar functionalities, but I will discuss on the pros and cons here. [4]

5.1 Application definition

Kubernetes can be deployed in combination of pods and microservices. Docker, whereas applications can be deployed as microservices.

5.2 Installation and Setup

Kubernetes is not very easy in terms of setup, but Google provides good documentation for the setup. It takes lot of time for the developer to get the setup done. On the other hand, Docker Swarm doesn't need much learning for the developers or devops and has easy process to setup and manage with the help of Command Line Interface (CLI). Overall, Docker wins here.

5.3 Monitoring and Logging

Both Kubernetes and Docker Swarm doesnt provide inbuilt support but extent offers logging and monitoring thru third party libraries. With Docker, DataDog, Reimann, Retrace and Sumo Logic can be used. With Kubernetes, Kibana and ElasticSearch can be used for Logging and Influx, Grafana, Heapster for Monitoring.

5.4 Size and Performance

Both can support 1000 node clusters in which each cluster can support up to 30,000 containers.

5.5 Conclusion

The best way to decide between the tools is probably to consider which one is more familiar and suits the existing software stack. Docker Swarm is a simple and easy solution to work with whereas Kubernetes is targeted those who requires support for complex applications.

6 SINGULARITY CONTAINER AND FEATURES

As per Singularity website "Singularity containers can be used to package entire scientific workflows, software and libraries, and even data. This means that you don't have to ask your cluster admin to install anything for you - you can put it in a Singularity container and run"

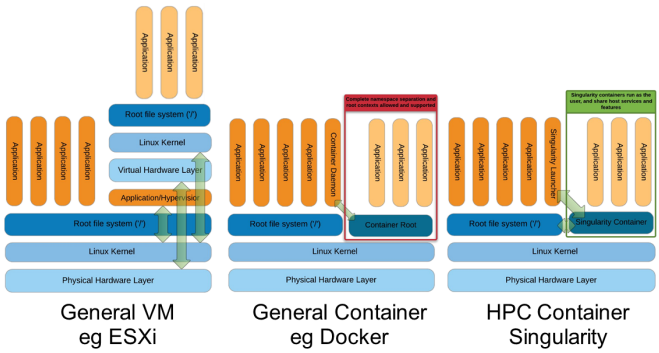


Figure 3: VM vs Docker vs Singularity [7].

Singularity containers is different from other containers due to the following aspects: **Application Portability** - Single Image File which contain all dependencies. Reproducibility, run cross platform and support legacy operating system and applications.
Docker Integration - It can convert Docker images into Singularity images easily
User Group Focus - It focuses on Scientific Application Users
Filesystem - Docker has no direct access to the host's filesystem except the directories are available cross mounted in the container. But for scientific computing, the access to host's files, data and libraries is needed. With Singularity process that is running as the user see the user's home directory and hence the user's environment is shared.
Security Model - It's the key aspect that it can be used with unprivileged permissions and doesn't require a separate daemon process. In which it highly useful in HPC workloads.
Docker containers run with root privileges in the container's operating system, which is not a problem if the container is running on a VM. But it's a security risk, if the container is running on a shared environment like university or super computer lab as the user can access the Docker daemon with the root access. Singularity solves the issue with creating and running the container with the user identity, so the permissions for the user will be same for both inside and outside the container.
[6]

ACKNOWLEDGMENTS

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

REFERENCES

[1] [n. d.]. Configuration-as-text. ([n. d.]). https://thenewstack.io/containers-container-orchestration/
[2] [n. d.]. Container Orchestration. ([n. d.]). https://dzone.com/articles/introduction-to-container-orchestration
[3] [n. d.]. Kubernetes. ([n. d.]). https://kubernetes.io/
[4] [n. d.]. kubernetes vs Docker Swarm. ([n. d.]). https://vexxhost.com/blog/kubernetes-vs-docker-swarm/
[5] [n. d.]. Platform9. ([n. d.]). https://platform9.com/blog/kubernetes-docker-swarm-compared/
[6] [n. d.]. Singularity. ([n. d.]). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426675/
[7] [n. d.]. vm docker singularity. ([n. d.]). https://platform9.com/blog/kubernetes-docker-swarm-compared/
[8] Provisioning. [n. d.]. . Technical Report. https://www.mongodb.com/containers-and-orchestration-explained

BigchainDB: A Big Database for the Blockchain?

Timothy A. Thompson
Indiana University Bloomington
School of Informatics, Computing, and Engineering
Bloomington, Indiana 47408
timathom@indiana.edu

ABSTRACT

Decentralized systems such as Bitcoin, the Interplanetary File System, and Ethereum have been designed with the intention of reengineering the architecture of online information networks, of minimizing exposure caused by central points of failure, and of creating new social models for the exchange of data—which is posited as a valuable asset in and of itself. Are these kinds of systems also able to support big data analytics and processing? If so, what stands to be gained by taking a blockchain-based approach to big data? Efforts to integrate blockchains into big data pipelines must inevitably address the tradeoff between security and scalability. BigchainDB is a new decentralized database framework that adds blockchain-based features, such as immutability and secure asset management, to traditional NoSQL distributed databases. Although it is still in the early stages of development, BigchainDB promises to make a significant contribution to the ways in which data is shared, processed, and managed at scale.

KEYWORDS

i523, HID340, Decentralization, Databases, NoSQL, Blockchains, BigchainDB

1 INTRODUCTION

Approaches to managing and processing big and complex data, such as the Lambda Architecture framework, have stressed the importance of treating data as an immutable asset [7]. In this view, data should never be updated, but only appended. In systems that allow data to be deleted or updated, big data only amplifies the surface of exposure to human error, and systems that conform to the standard relational database model of incremental updates become increasingly brittle as the scale of data increases [7]. Inadvertent deletions can trigger a cascade of data loss and system disruption that can be particularly difficult and costly to recover from. Even those who have criticized the specifics of the Lambda Architecture model (which proposes a complex internal division, within big data systems, between a batch layer and a realtime layer) agree that data immutability is an important foundation for building massively scalable platforms [6].

In decentralized, blockchain-based systems such as Bitcoin, immutability takes on an even more critical role. Without immutability and concomitant mechanisms such as Merkle tree hashing, it would not be possible to verify Bitcoin transactions for authenticity, nor would it be possible to maintain the "trustless" nature of the network—which is what allows decentralization itself to succeed [1]. In addition to Bitcoin, the emerging decentralized data ecosystem currently comprises platforms for computation (Ethereum) and file storage (Interplanetary File System—IPFS), but database software for managing metadata about assets is still lacking. BigchainDB is a databse solution that has been designed to fill this niche [8, 11].

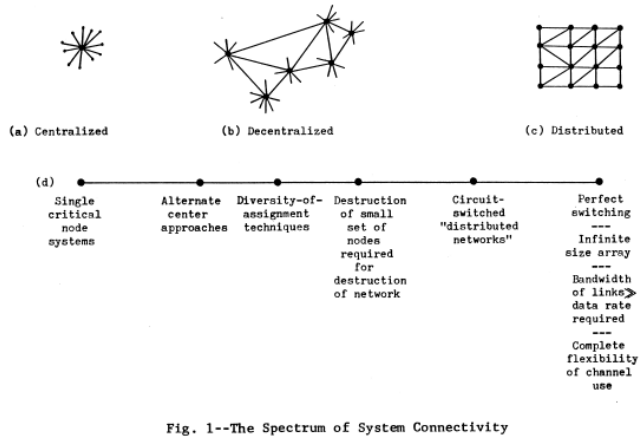


Figure 1: Baran's centralized–distributed network continuum [2]

2 DISTRIBUTED VERSUS DECENTRALIZED

Once it has been elevated to a core piece of system architecture, data immutability can become either a burden or an opportunity. The Lambda Architecture model leverages immutability within the context of an internally distributed environment, using storage solutions such as the Hadoop Distributed File System (HDFS) for processing on the batch layer [7]. Distributed systems are not the same as decentralized systems, however, and the terminology itself can be misleading, as illustrated by Baran's often-reproduced work depicting the continuum between centralized and distributed networks (shown in Figure 1). In Baran's model, decentralized networks are vulnerable to attack because of their reliance on hubs, whereas distributed networks are more durable because they employ a resilient grid-like structure [2]. In the context of the current discussion, "decentralized" systems such as Bitcoin are in fact exemplars of distributed models of connectivity. Distributed file systems such as HDFS, by contrast, may exist within highly centralized platforms or services. Here, the term decentralized will be used to refer to systems that embrace distributed models of organization both internally and externally.

2.1 Tradeoffs between Security and Scalability

Bitcoin's high level of security and resistance to attack are appealing features to engineers concerned with issues of data integrity. However, the Bitcoin network and blockchain are currently not

equipped to manage big data [4]. Compared to commercial financial transaction processors, which are capable of processing thousands of transactions per second, Bitcoin's computationally intensive "Proof of Work" model limits the network's throughput to a maximum of about 7 transactions per second [4]. Recommendations for improving the scalability of distributed ledgers such as Bitcoin range from adjusting system parameters (for example, maximum block size) to sharding the transaction validation layer in order to take advantage of parallel processing; however, the need for community approval and adoption of scalability solutions means that changes will take time to be implemented [4].

2.2 Blockchains for Big Data

The Bitcoin blockchain could itself be viewed as a globally distributed database, albeit not a particularly efficient or effective one. What is the primary benefit, then, of attempting to bring blockchain-based technologies to bear on big data? The potential value of integrating blockchains into big data pipelines is extrinsic rather than intrinsic: blockchains do not enable new analytical frameworks or algorithms per se, but rather promise to revolutionize the economies of exchange that determine the value and availability of big data [9]. Nevertheless, in order for big data pipelines to be "blockchainified," the issue of scalability must still be addressed. Because a systematic reengineering of blockchain systems such as Bitcoin seems unlikely in the near term, an alternative approach would be to modify existing big data systems by wrapping them with a blockchain layer. The latter approach is the one that has been adopted by the designers of BigchainDB [11].

3 BIGCHAINDB

BigchainDB is a database framework designed to extend the capabilities of existing NoSQL databases. To date, development of BigchainDB has focused on integration with two systems: RethinkDB (https://www.rethinkdb.com/) and MongoDB (https://www.mongodb.com/), although only the latter is recommended for usage as a production server [3]. Adding a blockchain layer to proven database systems allows developers to obviate or offload many of the problems with scalability and throughput currently faced by Bitcoin, for example. The BigchainDB whitepaper specifies three areas of innovation in which blockchain characteristics have been added as extensions to RethinkDB and MongoDB: decentralized control, immutability, and the creation and transfer of digital assets [11].

3.1 Decentralized control

A primary goal of BigchainDB's designers was to support seamless integration with the emerging ecosystem of "trustless decentralization" [11]. In order for a trustless approach to be viable, solutions to traditional distributed database problems such as consensus (how are conflicts among database nodes resolved?) must be implemented. In this regard, the BigchainDB whitepaper discusses three areas of concern: benign faults, Byzantine faults, and Sybil attacks [11]. To address benign faults (for instance, those caused by hardware failure), the default consensus protocol of the underlying database system is relied upon [11]. Although no claims are made for full Byzantine fault tolerance, the whitepaper goes into considerable detail regarding mechanisms for voting on and validating distributed transactions. Finally, Sybil attacks (in which a bad actor perpetrates a hostile takeover of a network) are addressed on the level of governance: the whitepaper describes a scenario in which BigchainDB nodes are deployed in a "federation with a high barrier of entry based on trust and reputation" [11].

3.2 Immutability

Data stored in a BigchainDB instance is treated as immutable and validated through a standard blockchain procedure of cryptographic hashing. Once a transaction has been approved by a majority of database nodes and fully validated, it is added to the system's internal blockchain, where is no longer exposed for updating or deletion. The approach to data management adopted by BigchainDB is known as CRAB (Create, Retrieve, Append, Burn)—in contrast to the CRUD (Create, Read, Update, Delete) model of traditional relational database systems [12]. Database assets that are "burned" are, in fact, not removed from persistent storage, but are simply reassigned to a randomly generated public key and effectively lost to the system [12].

3.3 Digital asset creation and transfer

At the core of BigchainDB is the framework that it provides for the secure transfer and tracking of assets in a chain of custody. Metadata surrogates can be created for both born-digital and real-world assets and assigned to individual users, identified by their public keys (from a system-generated public-private key pair). Assignment (that is, ownership) of metadata for assets can then be transferred from one user to another. The ability to make and authenticate claims about ownership is particularly valuable for content creators who wish to assert, and profit from, their intellectual property rights. Original development of BigchainDB was driven by the company ascribe.io, which provides artists with a platform for recording "art attribution, transfer, consignment and loan records" related to their work [5].

4 CONCLUSION

As a big data management system, BigchainDB promises to make it possible to attach "audit trails" to datasets, adding to their value by confirming their trustworthiness and integrity [9]. It also has the potential to serve as a decentralized clearinghouse for data sharing. Currently, access to data—including openly licensed data—is controlled by central service providers, in part to bolster assurances of the data's authenticity or accuracy. If data, instead, were stored in a decentralized blockchain database, it could be "collectively controlled by a public ecosystem" as an open marketplace [9]. Concerns about competitive advantage would be minimized because all parties would have a stake in ensuring the success of this global data exchange, and none would have the ability to control it directly or attempt to exclude other stakeholders from the opportunity to participate and profit.

Breaking down silos and providing incentives for data sharing also stand to benefit fields such as machine learning and artificial intelligence [10]. One lesson of the big data revolution was that

the accuracy of models could improve if enough data became available for analysis and training—and that simpler algorithms could often outperform more complex ones if the scale of data were large enough [10]. New possibilities for sharing data through decentralized, secure systems could now make it possible to leverage the scale of big data on an even bigger scale [10].

ACKNOWLEDGMENTS

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

REFERENCES

[1] A.M. Antonopoulos. 2017. *Mastering Bitcoin: Programming the Open Blockchain* (2 ed.). O'Reilly, Sebastopol, CA, United States.

[2] P. Baran. 1964. *On Distributed Communications: V. History, Alternative Approaches, and Comparisons.* Technical Report RM-3097-PR. RAND Corporation, Santa Monica, CA, United States. 51 pages. https://www.rand.org/pubs/research_memoranda/RM3097.html accessed 2017.

[3] BigchainDB Contributors. 2017. BigchainDB Documentation: Revision 3b33cdb1. Read the Docs. (2017). https://docs.bigchaindb.com/en/latest/index.html accessed 2017.

[4] K. Croman, C. Decker, I. Eyal, A.E. Gencer, A. Juels, A. Kosba, A. Miller, P. Saxena, E. Shi, E. Gün Sirer, D. Song, and R. Wattenhofer. 2016. On Scaling Decentralized Blockchains. In *Financial Cryptography and Data Security: FC 2016 International Workshops, BITCOIN, VOTING, and WAHC, Christ Church, Barbados, February 26, 2016, Revised Selected Papers*, J. Clark, S. Meiklejohn, P.Y.A. Ryan, D. Wallach, M. Brenner, and K. Rohloff (Eds.). Springer, Berlin, Germany, 106–125. https://doi.org/10.1007/978-3-662-53357-4_8

[5] BigchainDB GmbH. 2017. *A BigchainDB Primer.* Technical Report. BigchainDB GmbH, Berlin, Germany. 9 pages. https://www.bigchaindb.com/whitepaper/bigchaindb-primer.pdf accessed 2017.

[6] J. Kreps. 2014. Questioning the Lambda Architecture. O'Reilly.com. (July 2014). https://www.oreilly.com/ideas/questioning-the-lambda-architecture accessed 2017.

[7] N. Marz and J. Warren. 2015. *Big Data: Principles and Best Practices of Scalable Real-Time Data Systems.* Manning, Shelter Island, NY, United States.

[8] T. McConaghy. 2015. Blockchain Infrastructure Landscape: A First Principles Framing. BigchainDB Blog. (July 2015). https://blog.bigchaindb.com/blockchain-infrastructure-landscape-a-first-principles-framing-92cc5549bafe accessed 2017.

[9] T. McConaghy. 2016. Blockchains for Big Data. BigchainDB Blog. (Nov. 2016). https://blog.bigchaindb.com/blockchains-for-big-data-from-data-audit-trails-to-a-universal-data-exchange-cf9956ec58ea accessed 2017.

[10] T. McConaghy. 2017. Blockchains for Artificial Intelligence. BigchainDB Blog. (Jan. 2017). https://blog.bigchaindb.com/blockchains-for-artificial-intelligence-ec63b0284984 accessed 2017.

[11] T. McConaghy, R. Marques, A. Müller, D. De Jonghe, T.T. McConaghy, G. McMullen, R. Henderson, S. Bellemare, and A. Granzotto. 2016. *BigchainDB: A Scalable Blockchain Database.* Technical Report. ascribe GmbH, Berlin, Germany. 66 pages. https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf accessed 2017.

[12] J. Pregelj. 2017. CRAB—Create. Retrieve. Append. Burn. BigchainDB Blog. (Oct. 2017). https://blog.bigchaindb.com/crab-create-retrieve-append-burn-b9f6d111f460 accessed 2017.

Big Data = Big Bias? The Fallibility of Big Data

Gabriel Jones
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
gabejone@indiana.edu

Mathew Millard
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
mdmillar@indiana.edu

ABSTRACT

Since its origins, Big Data has promised to revolutionize the world. Scholars have wisely noted that it represents a paradigmatic shift from conventional norms of data, but the public has latched onto provocative but unrealistic narratives that deify Big Data as omniscient, infallible, and impervious to bias. Confiding in such narratives diminishes the integrity of credible science and poses serious ethical challenges, but these challenges are more likely overlooked because the problematic narratives seem to reject the need for ethical discussion. The authors argue that such blind optimism will cause irreversible damage to society if left unchecked. First, we debunk the fallacious narratives people tend to tell about Big Data, offering a more realistic discussion of its merits and its limitations. We then explore how analytical or algorithmic bias and sampling bias, two problems that statisticians have faced since long before the onset of Big Data, present pitfalls for deriving knowledge from data. We examine how the ethical implications of these pitfalls can cause serious damage in society. We conclude that effective, credible, and ethically sound Big Data analysis must obey the principles of transparency, clear and appropriate objective definition, and self-correcting feedback mechanisms.

KEYWORDS

i523, hid104, hid226, Big Data, Ethics, Algorithmic Bias, Sample Bias

1 INTRODUCTION: FALLACIOUS NARRATIVES ABOUT BIG DATA

In 2008, *Wired.com's* Chris Anderson wrote an article that captures the general optimism with which people conceptualize Big Data. The article, with its self-explanatory title "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", argues that Big Data provides such a complete, infallible view into reality that we no longer need conventional methods of scientific inquiry but need only to look at what the data tell us. According to Anderson, "With enough data, the numbers speak for themselves"[1]. This fervorous optimism was further extended in a 2013 book by Mayer-Schonberger and Cukier titled *Big Data* where authors assert that Big Data is synonymous with all data. In the past, researchers could only look at samples of data with limited scope, but Big Data, the authors claim, represents not a sample but a complete set[10]. A dataset of Twitter posts is viewed as synonymous with a complete, unbiased set of all of society's thoughts. By analyzing such a dataset, they conclude that they can confidently answer any question about how all of society thinks and behaves[9].

Cheerleaders for Big Data, such as Anderson, Mayer-Schonberger, and Cukier to make five exciting but yet flatly incorrect claims: that bigger is always better; that data analysis produces indisputably accurate results; that every data point can be studied, eliminating the need for archaic statistical sampling techniques; that studying causation is no longer needed since correlational patterns tell us all we need to know; and that scientific and statistical models are obsolete, since Big Data is itself sufficient. They have tended to extrapolate from the early success of the Google Flu Trends which at the time successfully embodied such grandiose, idealistic views. The Google Flu Trends project employed a theory-free set of algorithms that studied search engine results to predict flu outbreaks faster and more accurately than the Center for Disease Control. Allowing the numbers to "speak for themselves", Google determined that the number of searches about the Flu were correlated with flu outbreaks, so they concluded that more searches could accurately predict a greater spread[9].

At first it worked brilliantly. But in February 2013, just a month before the *n = all* proposition was published in *Big Data*, it made headlines for failing miserably, overestimating actual trends in 2013 by over 140 percent, leading Google to humbly terminate the program. The overconfidence of such an enormous dataset, viewed as a complete representation of reality free of gaps or inconsistencies, blinded them to its inherent flaws. For one, searches involving the term *influenza* are hardly an unbiased determinant of flu prevalence. They committed a classic statistical mistake by failing to consider confounding variables: the other reasons why people might search for the word *influenza*. Rather than adapting their model to fit changing patterns in the data, they assumed that the numbers could speak for themselves[9].

But blind proponents of Big Data bury the Google Flu Trends fiasco as just one not particularly convincing counterexample, giving superficial explanations that do not challenge Big Data's position as an infallible deity. In reality, such failure is the rule rather than the exception. Even Gartner, a company publicly known for pushing the importance of Big Data, estimated that 60 percent of Big Data projects would fail[8]. But it's not just a matter of occasional success or failure; many people in all disciplines misunderstand the nature of Big Data and therefore have unrealistic expectations. The narrative of the Target coupon case shows that society still regards the potential of Big Data as omniscient even if its execution is occasionally flawed. The story is narrated somewhat as follows.

In 2012, Target had collected enough purchasing data about pregnant women that they determined a particular high school girl was pregnant. When coupons for baby care items mixed in with general coupons started showing up in the mail, the father angrily visited the store manager to complain, suggesting that the store was encouraging teen pregnancy. The manager understood his frustration and called twice to apologize, but on the second call, the father's mood was different. The father offered his own apology

because Target was right. His daughter was pregnant, and Target's Big Data analytics managed to discover this before him[6].

While such a rose-colored narration fits well within the aforementioned grandiose conceptions of Big Data, a closer look shows that this successful case is overblown. While the anecdote seems to prove that Target's algorithms are infallibly accurate – that everyone receiving baby care coupons is pregnant – this is very unlikely. While the popular account suggests that Target mixes in coupons targeted towards pregnant women with other coupons to avoid spooking such women about their algorithmic accuracy, a much more credible explanation is that many women see mixed advertisements precisely because Target is unsure which ones actually are pregnant[9]. Even women who Target does suspect are pregnant have shopping interests outside of baby care items. While the algorithms help not to waste money by sending the coupons to, say, a single male adult living alone, they hardly indicate any reliable accuracy of pregnancy prediction. Of course, this is an empirical question that could be answered by researching how often pregnancy-targeted ads are sent to pregnant women versus those who aren't. But without having a methodologically sound study prove consistent accuracy, it's unwise to extrapolate from the anecdote and assume that Big Data done right is omniscient.

Critiquing the dominant reading of the Target case is not meant to suggest that Big Data has no value. Afterall, Target likely improved the efficiency of targeted advertising through Big Data by more accurately segmenting those who *might* be pregnant. But the important thing to keep in mind is that ultimately, models of the world and the data that feed them are imperfect. Models reflect the biases of those who create them, and data reflect biases inherent in sampling methods, time periods, and society in general. Cathy O'Neal, a former professor and Wall Street algorithm specialist with a mathematics degree from Harvard, observes that any model of the world "begins with a hunch, an instinct about a deeper logic beneath the surface of things"[12]. Human potential for bias and faulty assumptions can creep in. Of course, hunches or working thesis provide a necessary part of the scientific method of inquiry. Human intuition can be useful, as long there exist mechanisms by which those hunches can be evaluated and revised when necessary[12].

Perhaps the most common example of successfully wielding insightful models is depicted by the movie *Moneyball*, based on a true story. Oakland A's General Manager Billy Beane hypothesized that conventional performance metrics were overrated whereas more obscure measures better predicted overall success. He worked with statistician Bill James to create models that helped Beane decide which players to acquire and which to let go. The once obscure method has become a staple of baseball analytics. According to O'Neal, the model works for three main reasons: it allows for transparent analysis; its objectives are clear and appropriately quantifiable; and it includes a self-correcting feedback mechanism of new inputs and outputs, allowing it to be honed and refined. Models go wrong when they lack these three healthy attributes: "the calculations are opaque; the objectives attempt to quantify that which perhaps should not be; and feedback loops, far from being self-correcting, serve only to reinforce faulty assumptions"[12].

But models are only one factor in determining the efficacy of Big Data analysis. Since the very nature of data analysis is to extrapolate from limited samples, not only must researchers realize that models include human bias, but data itself is imperfect. It's true that data never lie. But it's false to assume they tell the truth. Data by themselves don't say anything; they simply are[4]. No matter how large and complex a dataset, it is always up to researchers to interpret the data to make meaningful claims. This is the essence of the scientific method that some want to reject.

2 ALGORITHMIC AND SAMPLE BIAS: THE THREATS THAT NEVER DISAPPEARED

Humans, as imperfect beings, should never assume that our creations are without flaw and bias. In many ways, mistakes and flawed thinking can trickle into the processes we come up with. This is the idea behind the fallibility of models created by humans with respect to algorithms used for handling Big Data. Some algorithms come with biases based on narrow thinking with a broad scope to cover. Other biases come from the assumption that the Big Data set being used is representative of the population when it really isn't. In any scenario, the creator is prone to introducing bias into any given algorithm, which can make it difficult to trust the results that the algorithm produces. With this in mind and considering the importance of specific findings, there is a lot at stake here. In some cases, lives can be changed for better or worse.

Sometimes algorithms, as models laden with the biases of their creators, can unintentionally manipulate readings of data in ways that reinforce false positives. But not all algorithms are wrong. In fact, machine learning shows us that often a well-written algorithm fed with good data can outperform human knowledge on everything from chess to medical diagnosis. But there's a problem with Big Data; it's inherently messy, complex, and distorted. Contrary to popular opinion that views it as a perfect representation of reality – recall the $n = all$ proposition – Big Data is a black box where typical issues with data quality hide themselves rather than disappearing. No matter how large or complex the dataset, the old adage still remains true: garbage in, garbage out.

The Literary Digest experienced the concept of garbage in, garbage out firsthand during the 1936 US presidential election, which pitted the Republican Alfred Landon against the wildly popular democrat Franklin D. Roosevelt. Roosevelt was particularly popular among the working class, the US majority, whereas Landon resonated well with the upper middle class and elites[9]. *The Literary Digest* Tried to predict the outcome of the election by sending out surveys to its own subscribers and by looking people up in phone and automobile registries. During the great depression, the people that owned phones, cars, and subscribed to the *The Literary Digest* tended to be more affluent and republican. After sending out 10 million ballots and receiving back nearly a fifth of them, they predicted that Alfred Landon would win with an astonishing 57 percent of the popular vote. They could not have been more wrong. Landon earned less than 40 percent of the popular vote, losing by a landslide[5]. This case has become the archetype example that data from a bias sample will lead to bias results. Increasing the volume of bad data only succeeds in producing a very precise incorrect conclusion, creating a false sense of confidence in something inherently wrong.

Although the *The Literary Digest* used lots of data, by definition their sample did not involve Big Data[10]. But if we reject the n

= *all* proposition, we can see that Big Data is still a sample and is therefore potentially vulnerable to sample bias. But while any statistically literate person can understand what went wrong with *The Literary Digest*, sample bias with Big Data is much more complicated and difficult to identify. For many people, random samples of social media data appear impervious to sample bias. Researchers conducting Twitter sentiment analyses often claim objectivity in representing the real world accurately, concluding that patterns observed in these vast, complex webs occur the same way offline. Despite the conflation of people and Twitter users, the two are not synonymous. Twitter users are by no means representative of the population. A Pew Research project in 2013 found that US-based Twitter users "were disproportionately young, urban or suburban, and black"[2]. To complicate things further, we cannot assume that Twitter data accurately represent how users behave because users and accounts are not a one-to-one relationship. Some accounts have multiple users, and some users own multiple accounts. Some accounts are just bots that automatically produce content, and some accounts are created and forgotten, going years without use. Furthermore, among active accounts, data are skewed by how some accounts dominate the discourse. Whereas some users post multiple times per day, others use the site only to view content. In fact, 40 percent of active users view content without making contributions, according to 2011 data from Twitter Inc[2]. The notions of what it means to be active, to participate, and to be a user require critical examination that's almost universally lacking.

The aforementioned examples highlight problems with available Twitter data, but there's also a problem with the integrity of available data. Twitter only makes a fraction of its data publicly available through its APIs. The supposed firehose of data theoretically contains all public tweets but explicitly excludes data that a user chooses to make private. Furthermore, theory does not match reality as the firehose lacks some publicly available tweets. Very few researchers get adequately full access. Research by Microsoft's Danah Boyd and Kate Crawford found that rather than a firehose, most have access to a "gardenhose (roughly 10 percent of public tweets), a spritzer (roughly 1 percent of public tweets)," or just select access through whitelist accounts[2]. Not only are protected data excluded, but data samples are not always randomized. So, a more reasonable description of Twitter data would say it takes a skewed sample of the real world population, further skewed by how users and bots create or do not create content, and then it limits the scope of the skewed data in an often opaque, arbitrary manner[2]. Is this data useful? Without a doubt. Is the data so perfect and infallible that we need not concern ourselves with basic principles of statistical and scientific credibility because "the numbers speak for themselves"[1]? Not even close.

If an algorithm could analyze a large, random sample of every word ever thought, spoken, or written by every human throughout their entire life, we could confidently believe that $n = all$ and make a sentiment analysis that accurately captures how people feel about a certain topic without regard for methods of scientific inquiry; the numbers would "speak for themselves"[1]. But we do not, and probably never will, have that kind of data. Twitter or other social media platforms are no substitute. While understanding the fallibility of Big Data is perhaps not as clear and straightforward as the *Literary Digest* case, society must be responsible by diligently scrutinizing data. To paraphrase loosely from world-renowned consultant Meta S. Brown, the biggest problem with data analysis will always be people failing to admit that data imperfections exist, failing to look for them, and refusing to do anything constructive about the ethical implications of these imperfections[3].

3 ETHICAL IMPLICATIONS OF ALGORITHMIC AND SAMPLE BIAS

As we've seen, the massive failure of the Google Flu Trends caused embarrassment and wasted Google's money. But the consequences they faced are relatively trivial, and given the company's history of learning from the past, they are probably a better company because of the failure. But when Big Data goes awry, the consequences are not always so trivial and localized. Big Data used unwisely has very serious, irreversible impacts upon society. Pervasive overconfidence can make it harder to acknowledge and confront such impacts until too late.

Society's current failure to address these issues is the topic of Cathy O'Neal's book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. She argues that these WMDs, referring to Big Data algorithms, have good intentions but often reinforce harmful stereotypes, especially of minorities and the poor, and become opaque models weidling arbitrary punishments. Through her work in the private sector, she has experienced numerous Big Data horror stories, and the book discusses several different failings of Big Data in various contexts.

One common issue associated with Big Data is the notion of self-fulfilling prophecy: the idea that expectations change reality to make it reflect the expectations. If police suspect African Americans to be more likely to commit crimes, they may patrol black neighborhoods more often and proactively hunt criminal activity. Increasing patrols increases the number of arrests, which provides justification to further increase patrols, causing more arrests, and so on. The prophecy that African Americans are more likely to commit crimes becomes adequately reinforced with their higher incarceration rates. But higher likelihood of arrest is not the same thing as being more likely to commit crimes[11].

It should be easy to see how the example of arrest rates is problematic, but somehow incorporating Big Data tends to make people fail to recognize the possibility of self-fulfilling prophecy. In fact, numerous police departments use algorithms that do just this, inadvertently instructing their officers to focus on areas with high concentrations of blacks. Crime prediction software that attempts to adjust police deployments according to anticipated patterns fail when they confuse more data with better data. Even though they attempt to prioritize violent and serious crime, data generated by relatively insignificant petty crimes, which occur far more often in poor and predominantly minority communities, can overwhelm the system, making it prejudice. Once the petty crime data enters a predictive model, more police deploy into those neighborhoods, and they are more likely to arrest people by their sheer presence and by the perceived threat that those people pose. The increased arrests justify the deployments in the first place[12].

But the danger does not end there. Once people are arrested by these inherently discriminatory processes, Big Data can work to keep them in prison for longer. This is usually not by intention

but by flaws in design. Recognizing how unconscious bias can affect sentencing decisions, courts in 24 US states have started to use computerized models to help assess the risk of recidivism, the likelihood of repeat offense. The models attempt to use Big Data to avoid a common, serious problem with human reasoning, and they certainly show some promise in this regard. But over reliance on the models can prove even worse than trusting potentially bias judges. "By attempting to quantify and nail down with precision what are at root messy human realities", the recidivism models shroud sentencing bias in a veil of unwarranted confidence and precise accuracy that disadvantages minorities by subjecting them to harsher prison sentences[12].

How does one quantify something as complex as the risk of recidivism? One popular model uses a lengthy questionnaire that attempts to pinpoint factors related to this risk. The questionnaire inquires about things such as previous police incidents. Given how much more often young black males get stopped by the police, partly because of the aforementioned self-fulfilling prophecy, such questions easily become a proxy for race, despite intentions to reduce this very prejudice. Other questions, such as whether or not the respondent's relatives or friends have criminal records, would be flagrant violations of court procedures and surely elicit objections from a defense attorneys if raised during a trial. But the opacity "of these complicated risk models shields them from proper scrutiny"[12]. Discriminatory police strategies feed into the recidivism models used to call for harsher sentencing, creating "a destructive and pernicious feedback loop"[12].

It is no secret that racial tension has become a dominant source of discussion when it comes to the american justice system. However, this issue is compounded with bias produced within the data itself as well. When there is a bias in how arrests are made based on the color of someone's skin, this bias feeds into an algorithm which opens up for more bias down the road. As more people of a given color are arrested and given harsher sentences, this data builds up in the system. The root of the cause may be human bias, but there is definitely a healthy amount of algorithmic bias that compounds and builds on the issue as most algorithms lack the ability to look beyond the face value of the data provided[7].

Big Data is, of course, not only used in attempts to more effectively dole out punishments. Facing international competition, Corporate America has latched onto its potential for increasing profits through more effective marketing, financial trading, and personnel decisions. With the prevalence of the internet, social media, and information literacy, Big Data presents an enormous opportunity for marketing personalization. Rather than targeting advertisement campaigns on broad, general audiences, Big Data can segment down to the individual level, targeting people based on their own personal data and patterns of behavior. However, this type of marketing is still a very inexact science and raises tricky ethical issues, including gender bias. Like racial bias, gender bias comes about in scenarios where profiling usually happens. For instance, advertising on the internet aims to reach its intended audience in order for businesses to sell products and make profits. Big Data and the statistical analysis involved might suggest that a certain gender has specific tendencies or lean on embedded societal stereotypes which cause some serious bias in an algorithm. One example might be a job opportunity being advertised. In this case, we want to say that either gender should be shown the advertisement a near equal amount, but we know from experience and outrage that this is not the case. It is almost staggering how it would favor the male population at times, especially when dealing with high paying jobs. Here, we also have a combination of Big Data and algorithmic bias working hand in hand to create biased results that ultimately lead to insult and faulty representation[3].

Beyond marketing, Big Data has found particular popularity among Wall Street investment firms, and for good reason. The ability to incorporate Big Data into decision making has tremendous potential for profitability. But the subprime mortgage crisis demonstrated how this can also have tremendous destructive potential. Financial models exhibited a particular bias, reinforcing the idea that what has worked in the past or what works currently will continue working indefinitely. But the sophisticated mathematical models lacked self-correcting feedback that could indicate inherent flaws. Since the models were driven by the market, if they led to maximum profits, they were considered infallible. Otherwise, why would the omniscient invisible hand of the market reward it? In hindsight we all recognize that betting on the subprime mortgage bubble was a losing proposition, yet the myopic reliance on the market proved disastrous in 2008. During the financial crisis, the algorithms used to assess securities risk became smoke screens. Their complex, mathematically intimidating design "camouflaged the true level of risk"[12]. The opaque models also lacked a healthy feedback mechanism that could have identified the problem[12]. The severity of the 2008 recession shows that companies are not only accountable for their own success and failure. Their misuse of Big Data had broad sweeping effects across the entire economy.

Perhaps it is reasonable to understand why companies might get carried away in a practice that, at least on the surface level, does not appear to affect humans directly. A trader working on the top floor of a Wall Street skyscraper might not see how the work of his mathematicians might hurt or harm average people. But Big Data also plays a role in ways that very clearly affect individuals, especially with the increasing popularity of integrating technology into personnel decisions. Since personnel decisions directly impact company performance, workforce management has become popular, particularly programs that promise to eliminate the guesswork from hiring by screening potential employees [12]. Many of these programs use personality tests to try and automate the hiring process; 60 percent to 70 percent of prospective employers, according to Deloitte Consulting.

Despite the optimism, such tests face the same problem as the recidivism surveys: they try unsuccessfully to quantify and precisely measure "what are at root messy human realities"[12] The high use of personality tests goes against research that consistently shows them to be poor predictors of future job performance. They don't provide this goal but rather an illusion of objectivity and simplicity. They generate raw data that get plugged into efficient algorithms and give clear answers, as opposed to the time consuming and obviously subjective process of human interviewing. Not only does this illusion cooly deceive companies, it leaves prospective employees disgruntled and confused by results from a opaque systems. Rejected employees don't know if they've been flagged or what caused them to be. The personality tests also lack important feedback mechanisms. There is no way to identify inherent errors

in the model and use those mistakes to refine the system[12]. Far too often, personality tests fail both the companies that use them and the prospective employees that get arbitrarily denied a chance.

In each of these cases, the story repeats itself where ethical issues that are normally fairly obvious become invisible when Big Data enters the picture. The argument is not that we should reject the positive potential of a reality that will only grow stronger with time. Rather, we should remain cognizant that a failure to adhere to basic principles of scientific credibility and ethical reasoning can affect people in unseen but deadly ways.

4 CONCLUSION

In the face of the copious amounts of new issues and problems we find around us when dealing with Big Data, there must be ways that we can hold Big Data and ourselves accountable. In order for Big Data to be the revolutionary force it promises to be, we must find ways to reduce bias and ultimately deal with ethical dilemmas in a proper manner. There are plenty of people around the world trying to solve these problems and progress is certainly being made. As humans, we will never be perfect, but understanding our imperfections and improving on our flaws is definitely a step in the right direction. Is there a way to catch our mistakes that we unwittingly make before we even know that we made them? Multiple cases studies suggest that the answer is a resounding yes: that we can make make algorithms which test for algorithmic bias. Such methods represent the future of Big Data; the idealized future will arrive when we successfully situate it within the broader context of data analysis in general, subjecting it to the same levels of scrutiny as we do for other types of data. We can simultaneously capitalize on Big Data's grand potential while avoiding ethical pitfalls when we successfully allow for transparent analysis; maintain clear, appropriately quantifiable objectives; and include feedback mechanisms that allow us to hone and refine the algorithms to produce objective results.

5 ACKNOWLEDGEMENTS

The author would like to thank Dr. Gregor von Laszewski and his teaching assistants for providing helpful feedback.

REFERENCES

[1] Chris Anderson. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Website. (June 2008). https://www.wired.com/2008/06/pb-theory/

[2] Danah Boyd and Kate Crawford. 2011. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. In *Six Provocations for Big Data*. https://ssrn.com/abstract=1926431

[3] Meta Brown. 2017. Math Isn't Biased, But Big Data Is. (AUG 2017). https://www.forbes.com/sites/metabrown/2017/08/30/math-isnt-biased-but-big-data-is/#2d6691dd4d56

[4] Kate Crawford. 2013. The Hidden Biases in Big Data. (April 2013). https://hbr.org/2013/04/the-hidden-biases-in-big-data

[5] Cynthia Crossen. 2006. Fiasco in 1936 Survey Brought 'Science' To Election Polling. (Oct. 2006). https://www.wsj.com/articles/SB115974322285279370

[6] Charles Duhigg. 2012. How Companies Learn Your Secrets. (Feb. 2012). http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=1&hp=&pagewanted=all

[7] Laurel Eckhouse. 2017. Big data may be reinforcing racial bias in the criminal justice system. (FEB 2017). https://www.washingtonpost.com/opinions/big-data-may-be-reinforcing-racial-bias-in-the-criminal-justice-system/2017/02/10/d63de518-ee3a-11e6-9973-c5efb7ccfb0d_story.html?utm_term=.0ee1409ec5c0#comments

[8] Laurence Goasduff. 2015. Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. (Sept. 2015). https://www.gartner.com/newsroom/id/3130017

[9] Tim Harford. 2014. Big data: are we making a big mistake? (March 2014). https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0

[10] Carl Lagoze. 2014. Big Data, data integrity, and the fracturing of the control zone. *Big Data and Society* 1, 2 (NO 2014), 1–11. https://doi.org/10.1177/2053951714558281

[11] Jasmine Liu. 2017. Big data and the creation of a self-fulfilling prophecy. (April 2017). https://www.stanforddaily.com/2017/04/05/big-data-and-the-creation-of-a-self-fulfilling-prophecy/

[12] Wharton. 2016. 'Rogue Algorithms' and the Dark Side of Big Data. (Sept. 2016). http://knowledge.wharton.upenn.edu/article/rogue-algorithms-dark-side-big-data/

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.

[2]

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 fi!? Central Intelligence Agency. 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. (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://dataconomy.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. (2016). (Accessed on 11/05/2017).

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 Analytics, Aviation Industry

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[7], 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 are 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[8].

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 [9].

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 [6].

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].

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 deman d[2].

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 [1]. 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 aircraftfis satellite communication system[11]. ACARS records values of different parameters during a event. An event is an action performed by the aircraft. The sensors mounted on the flightsfi 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 [2]. This portion of data sources generates the data at increasing rates due to the evolution of customerfis 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[12], 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[5]. 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[3].

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[3]. 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[10].

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 [4], 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].

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[2], 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 ACKNOWLEDGEMENTS

The author would like to thank Professor Gregor von Laszewski for providing all the help for this paper.

7 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.

REFERENCES

[1] A. M. Chandramohan, D. Mylaraswamy, B. Xu, and P. Dietrich. 2014. Big Data Infrastructure for Aviation Data Analytics. In *2014 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. 1–6. https://doi.org/10.1109/CCEM.2014.7015483

[2] EXASTAX. 2017. How Airlines are Using Big Data. EXASTAX. (01 2017). https://www.exastax.com/big-data/how-airlines-are-using-big-data/

[3] Maamar Ferkoun. 2015. Cloud computing helps airline industry soar. IBM. (02 2015). https://www.ibm.com/blogs/cloud-computing/2015/02/cloud-computing-helps-airline-industry-soar/
[4] Matthew Finnegan. 2013. Boeing 787s to create half a terabyte of data per flight, says Virgin Atlantic. COMPUTERWORLDUK. (03 2013). https://www.computerworlduk.com/data/boeing-787s-create-half-terabyte-of-data-per-flight-says-virgin-atlantic-3433595/
[5] T. Larsen. 2013. Cross-platform aviation analytics using big-data methods. In *2013 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. 1–9. https://doi.org/10.1109/ICNSurv.2013.6548579
[6] S. Li, Y. Yang, L. Yang, H. Su, G. Zhang, and J. Wang. 2017. Civil Aircraft Big Data Platform. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. 328–333. https://doi.org/10.1109/ICSC.2017.51
[7] Sebastien Maire and Chris Spafford. 2017. The Data Science Revolution That's Transforming Aviation. Forbes. (06 2017). https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/
[8] Maria Miller. 2017. How Data & IoT Technology are Changing The Way We Travel. (JUN 2017). http://dataconomy.com/2017/06/data-changing-way-travel/
[9] Eric Smalley. 2012. NASA Applies Text Analytics to Airline Safety. DataInformed. (07 2012). http://data-informed.com/nasa-applies-text-analytics-to-airline-safety/
[10] Air Vault. 2014. Virgin America Selects AirVault Cloud-Computing Service for Aircraft Maintenance Records Management. AirVault. (05 2014). http://www.airvault.com/news/virgin-america-selects-air-vault-cloud-computing-service-for-aircraft-maintenance-records-management/
[11] Wikipedia. [n. d.]. Aircraft communications addressing and reporting system. Wikipedia. ([n. d.]). https://en.wikipedia.org/wiki/Aircraft_communications_addressing_and_reporting_system
[12] Wikipedia. 2017. Cloud computing. Wikipedia. (2017). https://en.wikipedia.org/wiki/Cloud_computing

Using Big Data to Predict the Impact of Driverless Vehicles on the Unemployment Rate in the US

Paul Marks
Indiana University
Online Student
Shepherdsville, Kentucky 40165
pcmarks@iu.edu

ABSTRACT

A future of driverless cars is coming and with it comes a change in some industries. There will be both positive and negative impacts to jobs in these industries. The most obvious negative impacts are to driver-centric jobs such as taxi and truck drivers. Other industries, such as those directly related to producing and maintaining driverless cars, will see an incline in the number of jobs available. Data analytics can and will be used to help predict specific impacts and their effect on the overall United States (US) employment rate.

KEYWORDS

i523, hid327, self-driving, employment rates, driverless, technology and unemployment

1 INTRODUCTION

Overall there are an estimated four millions jobs[14] providing direct income to individuals and families which are based on a person driving a vehicle. The vehicle itself may vary, cars, short and long-haul trucks, and buses to name a few, but they all require driver today. That requirement, a human driver, will no longer be necessary in the future. Driverless vehicle technology is moving forward at fast pace. As the technology gets better and cheaper its ability to alter the employment scene in the US increases. "When autonomous vehicle saturation peaks, U.S. drivers could see job losses at a rate of 25,000 a month, or 300,000 a year."[1]

While most people think of negative impacts to employment due to driverless cars, less driving based jobs, there will also be positive impacts where jobs will be created. In order for the US to best meet the change in job demand, it must first be able to understand the impacts. Big data analytics can help to understand what these impacts may be, who they affect, where the impacts will be seen the most, and provide an opportunity for those impacted to prepare. This can be aided by data analytic models which can forecast the impact to industries in future years. The impact can then be leveraged more locally by breaking down the numbers state by state and city by city. Job losses will have an economic impact to the local economy while new jobs will have a positive impact. However, these changes may not generally occur in the same geographic area. Job losses and new jobs will also impact different demographics of workers.

2 JOB LOSSES

2.1 Passenger Transportation

One job category which will directly impacted by driverless vehicles is the transportation of people from place to place. Those providing on demand services such as taxis, ride hailing services (Uber, Lyft), and personal chauffeurs employed over 305,100 people with an average pay of $24,300 across the US in 2016[11]. Another 687,200 people were employed as bus drivers with an average pay of $31,920[7]. Combined this is almost one million jobs. These jobs tend to be more concentrated in larger cities. Once driverless vehicles are approved and people become more accepting of them these jobs will begin to disappear. The adoption rate could be faster than anticipated because many people may prefer driverless over a human driver for safety reasons. Not only is it expected that driverless vehicles will be much safer than those driven by people, they would never be approved if they are not, but some people are leery of getting into a vehicle with someone they do not know.

2.2 Goods Transportation

Aside from transporting people from place to place, vehicles are used to transport goods from place to place. In 2016 the heavy and tractor-trailer transportation industry employed 1,871,700 people directly driving the trucks with an average salary of $41,340[10]. In addition delivery truck drivers employed another 1,421,400 people with an average wage of $28,390[8].

Combining passenger and goods transportation accounted for almost 4.3 million jobs in the United States. The total wages earned by them was almost $150 billion. The average employment in 2016 across all twelve months was 144.3 million[9]. Comparing the employment numbers shows that approximately 2.97 percent of jobs in the United States rely directly on the ability of a person to perform the function of driving a vehicle. These jobs will be at an increasing threat from driverless vehicles once they are permitted on the roads. As the jobs are eliminated, so is revenue to federal, state, and local governments from the reduction of the $150B in wages.

2.3 Non-Driving

The impact of driverless vehicles does not only affect jobs which require a driver. Managing a team of drivers requires support from recruiting to human resources to managers. Many drivers eat at least one meal on the go so a reduction in drivers affects businesses such as restaurants, delis, and coffee shops. The ability of driverless vehicles to avoid accidents better than human drivers, some estimates up to 90 percent fewer accidents will mean less need for the

current 445,000 auto repair jobs[4]. It has to be assumed that driverless vehicles will be programmed to obey all traffic regulations so as the percentage of human drivers declines the need for vehicle enforcement and related court positions will decrease. The loss of driver income will decrease the amount of money to governments who would have to cut back on the number of people they employ. "Other peripherally-impacted jobs could include street meter maids, parking lot attendants, gas station attendants, rental car agencies, and more."[4] All of these items will factor into the impact of driverless vehicles on employment. To build a proper analysis of true, overall impact of driverless vehicles a model must include all these factors or it will not account for employment changes in indirect job markets.

3 NEW OPPORTUNITIES

Not all news about the anticipated proliferation of driverless vehicles is bad. Many industries will expand and new ones will start up. There are many companies investing in the hardware and software needed to make the technology viable. Driverless vehicles themselves will provide an opportunity for small businesses and entrepreneurs. If someone does not have to drive the vehicle anymore they will have time to do something else. Easy answers such as watching television or a movie, using their smart phone, or a computer are possibilities. However there are many untapped ideas yet to come for someone who sees the inside of a vehicle as an open space to be transformed into something different. It has been estimated that driverless vehicles "could add as much as $2 trillion to the US economy alone by 2050."[6]

4 PREDICTING THE IMPACTS

So how can big data analytics help to predict the impact of these changes on employment in the United States? The key will be creating good predictive models. The answer is not as simple as estimating the number of jobs which will decrease and comparing it to the number of jobs expected to increase. Employment and unemployment rates are much more complex to predict. "If a firm lays off 1000 workers, only a fraction will enter the ranks of the unemployed"[12]. This may seem surprising to some, but it is explained by the dynamics which need to be modeled. One of the premises of employment models is that "workers live forever, spending their lives moving between unemployment and employment"[13]. This is not saying that each individual worker lives forever, but that as a whole workers need to work and will actively seek to be employed.

A big data analysis must also take into account the demographics of the workers. 4.23 percent of African American workers are employed in a driving occupation and 3.25 percent are of Hispanic descent. The top five states which employ drivers are California, Texas, New York, Florida, and Illinois. These jobs pay on the average more than non-driving jobs for similarly skilled people[14]. This means that the jobs being lost are not easily replaced with jobs of the same pay. So even for those who are able to find other work it means that their standard of living will be reduced.

The further complicate predicting the impact of driverless vehicles is the difference in jobs skills needed. Many of the jobs lost are lower skill jobs; they do not require degrees. Many of the jobs being generated by the driverless car industry will be higher skills such as analyzing all the data which will be generated by the vehicles and their passengers. (Wired) That means the model must take into account "the match between the searchers' characteristics and those of the available jobs"[12]. Another issue with the impact on employment is that it will not be homogeneous throughout the United States. Studies show that "job seekers in depressed areas may not be able or willing to relocate to areas with better job prospects"[5]. This means that jobs lost in one area are not filled by those who lost their jobs if the replacement job is not in the same area.

5 USING THE ANALYSIS TO CURTAIL THE IMPACT

One of the models used in employment calculations is the Aggregate Demand/Aggregate Supply model. Aside from demand and supply it takes into account "a wide array of economic events and policy decisions" [3]. This is one important aspect of any model used to help predict the impact of driverless cars on the economy; the fact that policy decisions are part of the equation. Driverless cars will not become a reality without federal, state, and local governments passing legislation around their use. In doing so the impact of driverless cars may be lessened or spread out over time.

The models can also be used to help minimize the impact on unemployment. In predicting the impact of the change in technology and what areas will be impacted the most, the government will be able to proactively take steps to train, retrain, or put other efforts into place to offset the impact. By doing this type of analysis proactively the impact of such changes can also be predicted. Providing the right information to people can reduce unemployment, shorten the time people are without jobs, and decrease the time that jobs go unfilled. However it takes data analysis to do this. Job openings and losses, must be examined with other data sources such as "changes in the distribution of jobs across industries and regions, shifts in the demographic characteristics of the work force, and other changes in the way labor markets operate"[2].

6 CONCLUSIONS

Driverless vehicles will become a reality in everyday life in the near future. The ability to move people and cargo from point to point without the need of a human driver will impact the millions of jobs which are based on driving. On the surface this does not seem like an issue which requires too much data analysis; it is easy to show that a loss in jobs affects employment rates. However when it comes to employment the US cannot afford to be passive. The solution requires input from many data sources covering not only the jobs, but geographic information and demographic data on those affected.

By leveraging big data analytics, models can be created which can predict what this impact is, what geographic locations will be impacted, and which population demographic will be affected the most. The results of such models may be stark and they may be surprising. What is important is to create data models which will be based on factual data so that it can be understood before the changes take place. By doing this, by leveraging data, a potential major impact to the United States employments rate can be seen in advance and prepared for.

This is the real power of using big data analytics to study the impact driverless cars will have on the workforce: knowing what is coming before it happens. This can allow for plans to be made to minimize the impacts and prepare those who could be affected the most. This can have a positive result not only at a national level, but specifically at an individual level for workers who will be displaced, but get help in preparing for the change which sets them up for another opportunity. This type of analysis is how advances in data technology can help society predict and prepare for change in a positive manner.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support of this topic. While the impact of driverless cars on the employment rate does not necessarily seem to be a big data issue on the surface, it became more and more an issue which big data can help with. This is one of the many uses of big data which will benefit society, the ability to help prepare for future issues.

REFERENCES

[1] Anita Balakrishnan. 2017. Self-driving cars could cost America's professional drivers up to 25,000 jobs a month, Goldman Sachs says. Online. (05 2017). https://www.cnbc.com/2017/05/22/goldman-sachs-analysis-of-autonomous-vehicle-job-loss.html

[2] Kelly A. Clark and Rosemary Hyson. 2001. New tools for labor market analysis: JOLTS. *Monthly Labor Review* 124, 12 (2001), 32–37. http://www.jstor.org/stable/41861552

[3] OpenStax College. 2015. How the AD/AS Model Incorporates Growth, Unemployment, and Inflation. Online. (09 2015). https://legacy.cnx.org/content/m57327/1.5/

[4] Joel Lee. 2015. Self Driving Cars Endanger Millions of American Jobs (And Thatfis Okay). Online. (06 2015). http://www.makeuseof.com/tag/self-driving-cars-endanger-millions-american-jobs-thats-okay/

[5] Ioana Marinescu and Roland Rathelot. 2016. *Mismatch Unemployment and the Geography of Job Search.* Working Paper 22672. National Bureau of Economic Research. https://doi.org/10.3386/w22672

[6] Aarian Marshall. 2017. ROBOCARS COULD ADD $7 TRILLION TO THE GLOBAL ECONOMY. Online. (06 2017). https://www.wired.com/2017/06/impact-of-autonomous-vehicles/

[7] United States Department of Labor. 2017. Bus Drivers. Online. (10 2017). https://www.bls.gov/ooh/transportation-and-material-moving/bus-drivers.htm

[8] United States Department of Labor. 2017. Delivery Truck Drivers and Driver/Sales Workers. Online. (10 2017). https://www.bls.gov/ooh/transportation-and-material-moving/delivery-truck-drivers-and-driver-sales-workers.htm

[9] United States Department of Labor. 2017. Employment, Hours, and Earnings from the Current Employment Statistics survey (National). Online. (11 2017). https://data.bls.gov/timeseries/CES0000000001

[10] United States Department of Labor. 2017. Heavy and Tractor-trailer Truck Drivers. Online. (10 2017). https://www.bls.gov/ooh/transportation-and-material-moving/heavy-and-tractor-trailer-truck-drivers.htm

[11] United States Department of Labor. 2017. Taxi Drivers, Ride-Hailing Drivers, and Chauffeurs. Online. (10 2017). https://www.bls.gov/ooh/transportation-and-material-moving/taxi-drivers-and-chauffeurs.htm

[12] Jeffery Parker. 2010. Economics 314 Coursebook. Online. (2010). http://www.reed.edu/economics/parker/s11/314/book/Ch14.pdf

[13] Thomas Sargent and John Stachurski. 2017. A Lake Model of Employment and Unemployment. Online. (2017). https://lectures.quantecon.org/jl/lake_model.html

[14] Andrew Simpson. 2017. 4 Million Driving Jobs at Risk from Autonomous Vehicles: Report. *Insurance Journal* March 27, 2017 (03 2017). https://www.insurancejournal.com/news/national/2017/03/27/445638.htm