Use Cases in Big Data Software and Analytics

Vol. 1, Fall 2017

Bloomington, Indiana

Wednesday 13th December, 2017,
12:59

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	Jones, Gabriel	What Separates Big Data from Lots of Data?
105	Lipe-Melton, Josh	This is my paper about data visualization in sports
106	Qiaoyi Liu	Big Data Analytics in Groceries Stores
107	Ni,Juan	DevOps in support of Big Data Applications and Analytics
109	Shiqi Shen	Big Data in Social Media
201	Arnav, Arnav	Big Data analytics and Edge Compting
202	Himani Bhatt	Big data analytics in Weather forecasting
203	Chandwani, Nisha	Big Data Analytics using Spark
204	Chaturvedi, Dhawal	Big Data Analytics and High Performance Computing
205	Chaudhary Mrunal L	Applications of Big Data in Fraud Detection in Insurance
208	Jyothi Pranavi,Devineni	Big Data and Deep Learning
209	Han, Wenxuan	Big Data Application in Web Search and Text Mining
210	Hotz, Nicholas	Natual Language Processing of Electronic Health Records
211	Khamkar, Ajinkya	Distributed environment for neural network
213	Yuchen Liu	Big Data and Speech Recognition
216	Mathew Millard	Big Data Analytics in Sports - Track and Field
224	Rawat, Neha	Big Data Applications in the Hospitality Sector
225	Schwartzer, Matthew	Creating Better Urban Environments with Optimized Public Bus Routes and Schedules
228	Swargam, Prashanth	Big data applications in Electric Power Distribution
229	ZhiCheng Zhu	Big Data and Machine Learning
231	Vegi, Karthik	Using Big Data for Fact Checking
232	Rahul velayutham	Big Data Analytics in Sports - Soccer
233	Wang, Jiaan	Big Data Applications in the Media and Entertainment Industry
234	Weixuan Wang	Big Data Analytics in Tourism Industry
235	Yujie Wu	Big Data in Recommendation System
236	Weipeng Yang	Big Data in MOOC

237	Tousif Ahmed	Big Data Analytics in Cyber Security and Threat Research
301	Gagan Arora	Big Data Analytics in Finance Industry
302	Sushant Athaley	Big Data Application in Restaurant Industry
304	Ricky Carmickle	Big Data and Astrophysics
305, 319	Andres Castro Benavides, Mani Kumar Kagita	Big Data Analytics for Municipal Waste Management
306	Murali Cheruvu	The Internet of Things and Big Data Analytics
309	Dubey, Lokesh	BigData Analytics using Apache Spark in Social Media
311	Matthew Durbin	Big Data and Healthcare
312	Neil Eliason	An Overview of Big Data Applications in Mental Health Treatment
313	Tiffany Fabianac	Big Data Platforms as a Service
315	Garner, Jeffry	Roles and Impact on Mobility Network Traffic in Big Data
316	Robert Gasiewicz	Big Data Analytics in Biometric Identity Management
305, 319	Andres Castro Benavides, Mani Kumar Kagita	Big Data Analytics for Municipal Waste Management
320	Elena Kirzhner	Big Data Analytics and Applications in Childbirth
323	Uma M Kugan	This is my paper about NoSQL Databases in support of Big Data Applications and Analytics
324	Ashok Kuppuraj	Big data and Analytics in Blockchain
325	J. Robert Langlois	Impact of Big Data on the Privacy of Mental Health Patients
326	Mohan Mahendrakar	Bigdadta in Clinical Trails
327	Paul Marks	Using Big Data to minimize Fraud, Waste, and Abuse (FWA) in United States Healthcare
328	Dhanya Mathew	Big Data Analysis in Finance Sector
329	Ashley Miller	Big Data Analytics in Higher Education Marketing
330	Janaki Mudvari Khatiwada	Big data in Improving Patient Care
331	Tyler Peterson	Big Data Applications in Population Health Management
332	Judy Phillips	Big Data Analytics in Agriculture
337, 333	Ashok Reddy Singam, Anil Ravi	Big Data and Artificial Intelligence solutions for In Home, Community and Territory Security
334	Peter Russell	AWS in support of Big Data Applications and Analytics
335	Sean M. Shiverick	Big Data Analytics, Data Mining, and Public Health Informatics: Using Data Mining of Social Media to Track Epidemics
336	Jordan Simmons	Recommendation Systems on the Web
337, 333	Ashok Reddy Singam, Anil Ravi	Big Data and Artificial Intelligence solutions for In Home, Community and Territory Security
338	Anand Sriramulu	Docker in support of Big Data Applications and Analytics
339	Hady Sylla	Big data application for treatment of breast cancer
340	Timothy A. Thompson	Big Data Analytics for Research Libraries and Archives
342	Udoyen, Nsikan	Big data analytics in college football (NCAA)
343	Borga Edionse Usifo	Big Data Applications in Self-Driving Cars
345	Ross Wood	Big Data Dangers Weaponizing Social Media
346	Zachary Meier	Big Data in Oceanography
347	Jeramy Townsley	Sociological Applications of Big Data

Big Data Analytics in Groceries Stores

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

ABSTRACT

This paper helps us understanding how big data is working in Groceries store and how Big Data helping their business.

KEYWORDS

i423, hid106, Data Science, Big Data Analytics, Cloud Computing,customer study

1 INTRODUCTION

Today, numerous market chains perform an assessment of their client/customers on a massive set of data, discovering experiences that assist them better includes customers and, thus, drive income. Discerning how to use big data is vital in an industry where profits are razor thin, and waste management is a broad issue. By gathering and evaluating customer data, grocery stores can sharpen their approach to everything from advertising exercises and pricing to product classification and customerfis benefit [2, 4]. With the appropriate analytical tool, grocery stores can unite various sources of data and get information progressively in real time, letting them precisely conjecture product demand, improve stock levels and turn-rates, and lessen waste of perishable products. The article will examine the importance of Big Data Analytics in Groceries stores, its relevancy, its use as a competitive advantage tool to attracting customers, in addition to determining customer demand. Grocer Loyalty program databases, rich with a point to point customer information, have been in presence for quite a long time, giving food merchants a clear preferred standpoint (for those that have utilized this information) contrasted with different retailers [3]. Grocers have had a head start beginning on using this information in better understanding shopper behaviors and shopping preference. Nonetheless, with the coming of new technological innovations, new contenders, new channels and the rise of a 'constantly-on' and 'time-starved' purchaser base with a bunch of advantageous shopping choices fi?! the grocery industry is presently trailing different retailers in the capacity to use these new 'huge' data sources to advance their investigative abilities from interactions to transaction[1]. Specifically, the development of new types of data sources fi?! big datafi?! offers a chance to Small to Medium Size grocer's equal opportunity to compete with big chains of supermarkets. The proceeding section highlights the relevance of the big data.

2 RELEVANCY OF BIG DATA ANALYTICS IN GROCERIES STORES

2.1 Increases the customer shopping experience

As per a current SHSFoodThink white paper "Are We Chain Obsessed?" 64% of customers said that the previous shopping experience is what makes them keep coming backfi!?not the items themselves [5]. By utilizing bits of knowledge received from the information transaction database, online networking, promotional activity, customers purchasing behavior, and client movement patterns, grocery stores can find a way to guarantee they are engaged with their customers that matter most. For instance, they can investigate customers shopping movement to enhance the layout of their store, or recognize attrition risks for clients who have not as of late bought staple things, similar to milk. In like manner, chains can construct item varieties demonstrated with the customer needs and purchase patterns in certain regions [1, 3, 5]. Regardless of whether it is through reconsidering store layout or furnishing store attended with mobile apps to better serve clients, analytics can enable grocers to change consumerfis expectations.

2.2 Restructure the Supply Chain

Grocery stores can likewise utilize analytic to investigate the production of their products, monitor production processes, and quality control, and improve straightforwardness with buyers about their sustenance production practices of foods [4]. Suppliers remain to profit from the evaluation also, with access to secure, customized content of information identified with performance sales of the product, stock, margins, and marketing effectiveness. Giving supplier an opportune profitable business knowledge that supports joint ventures, drives performance, and decreases waste products

2.3 Build Superior Marketing Programs

Loyalty programs furnish grocery merchants with an abundance of data to enable them to distinguish client segments and precisely characterize item preferences. By joining this information with different data sourcesfi!?like healthful patterns, favored technique for accepting marketing promotion, customer movement patterns, and weather-related eventfi!?grocery merchants can concentrate on enhancing, and derive income from, the general shopping experience [5]. For instance, grocery retailers can utilize analytics to customize the advancements they offer to clients given what they are well on the way to buy. They can likewise time advancements fittingly, and offer codes to customers who often as possible buy certain things.

2.4 Improves HR Strategies

Supermarket stores utilize analytics to manage work-related decisions. Information freely accessible through online networking accounts and different means can be examined in conjunction with a grocer's internal information to direct decision identified with selection and recruitment, employee termination, and performance management and advancements [2]. For example, an investigation of late action on LinkedIn can reveal insight into which representatives are destined to leave an organization. Grocery merchants can likewise break down information to control the advancement approaches that will build workforce performance. For example, they could explore different avenues regarding organizing a social gathering for representatives at a subset of their stores, and analyze information on profitability, morale, and turnover in the preceding months [3]. They may find that the gathering information prompted a more positive workplace where workers feel more noteworthy engagement at work, and soon after that, they could roll the strategy out to different stores.

2.5 Using big data for competitive advantage and attracting customers

Numerous grocery stores have been utilizing transaction and client information for a considerable length of time, despite the fact that many still have not completely used all that can be proficient with these types of information. For Small to Medium Sized grocery merchants, many have swung to subcontracted point solutions because of an absence of available analytics assets and potential framework investment required [2, 5]. The issue with point solutions recently is that fi?! they independently work out for a particular business section and the evaluation is cookie cutter. In this way, the 'information' is not coordinated and hard if not difficult to give an all-encompassing picture of client conduct overall touch focuses for instance. Nor are the investigations offering a cross-functional observation that is pertinent to all business partners as far as driving differentiation in the commercial center in promoting, advertising, store operations and supply chain. As far as utilizing 'new' data sources, for example, mobile, social and text, the industry is particularly occupied with a discovery' phase of investigation with an assortment of center sections, testing and figuring out how to extricate an incentive from these rich new sources of information. There are two common paths grocery merchants takes with little respect of the 'size' of the organization: to start with is Strategic Commitment, in which there is C-level (hierarchical) commitment making the venture in the assets to get the majority of the in-house data and evaluated it [3]. Presently like never before, information, analytics, and IP are seen as vital resources and competitive discriminators. The other is Business Discovery; in which grocery merchants outsource to an Analytics as a Service firm to use internal and external information. Performing analytics speeds the construction of business advantages creating new users case and helps catch 'quick wins' before making resource commitment to technological innovation and human capital in advance [2]. In view of progress, and a wit, trusted stakeholder willing to share the techniques and explanatory models, can assist grocery merchants to proceed with an outsourced administrations supplier or relocate the data, analytics in addition to IP in-house.

3 RECOMMENDATIONS

3.1 Real-time insight on product demand

Nowadays, retailers can get to information on item demand levels instantly on a chain of stores. Nevertheless, numerous merchants are still in the earliest stages in regards to evaluating and monetizing the huge amount accessible data [1]. This prompts stocking deficits, for example, evaluating item demanded based exclusively on past historical information. It can likewise convey about wrong promoting endeavors: If a customer purchased ketchup on Saturday, an email coupon for it on Sunday is not well planned and make little sense to the shopper. This is the place data from store loyalty programs in addition to credit card sales can prove to be useful. Its data can be utilized to define needs of the customers in future. For example, grocery merchants can use data analytics to decide how regularly customers purchase sugar, flavors, or different items, and after that send every family unit coupons given their propensity to buy [5].

3.2 Enhancing in-store stock management

Perishable basic supplies, for example, dairy, meat, and fish call for precise stock administration, regularly on an hourly premise. Client analytics and prediction tools can enable grocery merchants to calibrate their inventory levels by assessing buyer purchasing behavior and requested products from various viewpoints and situations [5]. For example, grocery retailers might need to screen cycles like when customers go for particular nourishment, purchasing patterns amid sales deals when storing activity peaks or seasonally inspired buys. As indicated by a report from Manthan, this methodology worked for U.K. food grocery merchant Waitrose: a deeper understanding of buyer purchasing behavior and demand outlines using cutting edge client analytics and predicting tools helped the store [2]. Concurrently, retailers can utilize these systems to all the more deftly change their stock levels and amplify high-buy products.

3.3 Leveraging Predictive Analytics

Amazon spearheaded item proposal engine: the "if you purchased that, you may like this" invention. This strategic changing web-based shopping feature mirrors the retailer's profound assessment of buyers' shopping basket. Proposal engine is intended to enable customers to find items they were not sorting out but rather would be interested in purchasing [3]. Today, general grocery merchants are progressively tapping the global innovation behind proposal engine: predictive analytics. This kind of assessment measures future patterns in light of present and past information, and it can enable stores to improve business. Information is driven, all-encompassing assessment of "purchasing triggers, for example, regularity, weather, stock, and advancements, is progressively informing grocery stores' product blend, marketing plans, and sales forecast [5]. Furnished with these information-driven tools, stores can better distinguish what items customers need today and what they will be demanding in future, and this learning will enable them to stay competitive for a considerable length of time to come.

4 CONCLUSION

Big data analytics is profound tool assisting grocery merchant establishing insightful information concerning the market structure and sales demand. With the appropriate analytical tool, grocery stores can unite various sources of data and get information progressively in real time, letting them precisely conjecture product demand, improve stock levels and turn-rates, and lessen waste of perishable products. Advancement in information technology is offering new means fi?!Big data analytics that Small to Medium Business such as grocery merchant can use to drive the products sales. Big as discussed previously, assist grocery merchant to increase their customer experience, restructure supply chain, create superior market programs, improve HR strategies, and creates them a competitive advantage.

ACKNOWLEDGMENTS

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

REFERENCES

[1] J. Aloysius, H. Hoehle, S. Goodarzi, and V. Venkatesh. 2016. Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. *Annals of Operations Research* (2016).

[2] Ban G-Y. 2014. Business analytics in the age of big data. *Business Strategy Review* 25, 3 (2014), 8–9.

[3] M. Ghesmoune, M. Lebbah, and H. Azzag. 2016. State-of-the-art on clustering data streams. *Big Data Analytics* 1, 1 (2016).

[4] A. Hussain and A. Roy. 2016. The emerging era of Big Data Analytics. *Big Data Analytics* 1, 1 (2016).

[5] Eric Siegel. 2013. *Predictive analytics: the power to predict who will click, buy, lie, or die.* Vol. 51. Wiley; 1 edition.

Big Data Applications in the Hospitality Sector

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

ABSTRACT

The rise of *Big Data* in the field of Hospitality though recent, is by no means temporary. The hotel industry is one which deals with millions of customers on a day-to-day basis and generates a plethora of customer data through such interactions. It is also the sector which depends the most on customer loyalty, and thus profits greatly through the analytical insights that Big Data has to offer. Keeping this is mind, hotels today, whether they are big chains or small independent establishments, are using data generated internally and on the web to develop strategies for better customer satisfaction, marketing effectiveness, yield management and operational efficiency.

KEYWORDS

i523, HID224, Marketing, Yield Management, Recommendation Systems, Data Warehousing, Data Mining

1 INTRODUCTION

Big data is often defined as "data that exceeds or is beyond the capabilities of the organization to store or analyze for accurate or timely decision making" [11]. It is characterized by features such as its volume, velocity and variety. Two other characteristics that have been recently added to these are veracity and volatility, referring to the uncertainty and dynamic nature of such data [11]. Despite the unstructured nature of such data, it presents us with a variety of opportunities which make it so appealing.

The hospitality sector too generates a huge amount of data in its day-to-day processes about its customers, operational processes such as electricity and water consumption and the daily revenue generated. Some of the questions that can be addressed using this data are - What is the country of origin of the customer? What are his/her particular preferences in terms of food or other amenities? What booking channel did they use? What was the time/season of booking? How is the performance of the hotel relative to the local market? What is the monthly energy consumption and other expenditures?[1].

Using the data generated internally by the administrative units and different departments, gathered externally from the web - from sites of aggregators such as Expedia and Trivago and from social networking sites such as Twitter, hotels can derive quite useful insights into the opportunities they can utilize and the challenges they should overcome.

2 THE ADVENT OF BUSINESS INTELLIGENCE

Business Intelligence has been a part of the Hospitality sector for some time now. Earlier though, it was used mainly in traditional revenue management systems to deal with duration of stays and promotional programs [6]. However, it was not developed well enough to provide additional insights into other areas of hotel management. The emergence of companies which worked as online booking platforms, such as Expedia and Travelocity, led way to a new form of data, which though unstructured, could be leveraged as a window into the customers' preferences. Few hotels such as Mariott, InterContinental Hotels, Hilton and Hyatt utilized these opportunities presented by business intelligence to get ahead in the game, but were not very successful due to insufficient planning and technical expertise, lack of executive support and wide-spread adoption [6].

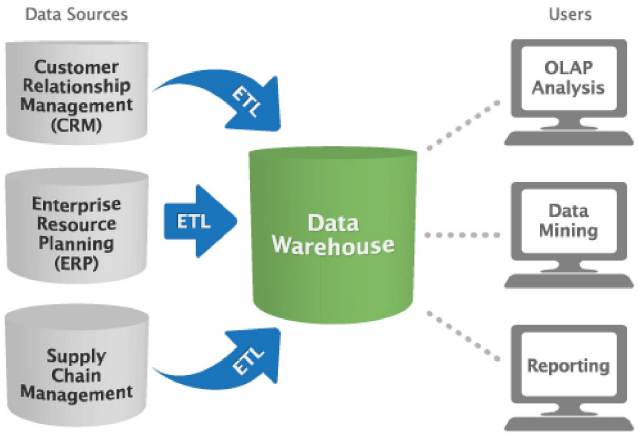

Figure 1: Hospitality Business Intelligence System [2]

3 BIG DATA AND HOSPITALITY

With the advancement of technology, the generation of data increased manifold. Researchers from UC Berkeley had estimated that "the world had produced about 1.5 billion gigabytes of information in 1999 and in a 2003 replication of the study found out that amount to have doubled in 3 years" [12].

The Hospitality sector too found data from a variety of sources - social media, review data, data from search engines like Google and other customer data sources [1]. As a result, business intelligence systems were made more structured, driven by advanced IT technologies and machine learning algorithms. Figure 1 shows a Hospitality Business Intelligence System

The vast amount of data now is used for a variety of tasks in the hotel management field. The most essential ones are - Customer Satisfaction and Marketing, Yield/Revenue Management and Operational Effectiveness.

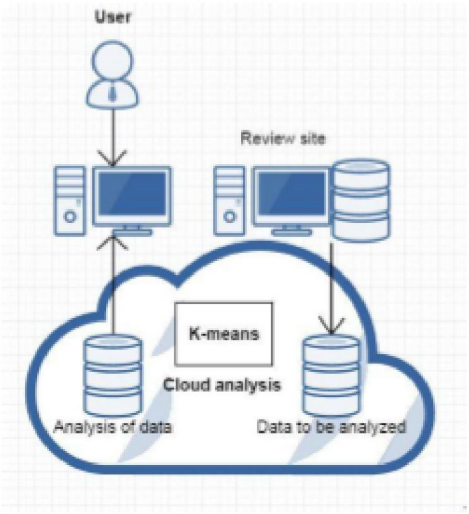

Figure 2: Structure of the Customer Response Evaluation System [4]

3.1 Customer Satisfaction and Marketing

Customer loyalty is the main driver behind the hospitality business. The use of big data analytics has worked towards providing hoteliers with insights about what their guests want. This information can be used by the hotels to improve existing customer satisfaction as well as develop marketing techniques to attract new customers. This helps convert "high-spending customers to repeat customers" and increases the hotel's profitability [8].

An excellent example of the use of analytics for customer satisfaction is the new system introduced by the US chain of hotels, Denihan Hospitality [7]. Using IBM analytics technology, they worked towards combining their internal customer and transactional data with the review data found on the web. This was used to implement various data-driven solutions regarding the quality of the rooms, bathrooms and other facilities. They even went ahead to create interactive dashboards and "putting analytics in the hands of the frontline hotel staff" who received real-time updates as to the requirements of their customers [7].

In order to implement the above hotel evaluation structure, one can use the services of WebCrawlers and cloud computing platforms like Hadoop. A similar system created by Ming-Shen Jian, Yi-Chi Fang, Yu-Kai Wang and Chih Cheng uses cloud technologies coupled with data mining algorithms to create a customer response and evaluation system [4]. The system uses Hadoop to implement a multi-node cluster on the cloud server, programs a WebCrawler to retrieve review data from websites, extracts the informative adjectives using MapReduce and a text segmentation system and gives the word count using Hadoop's WordCount program. Weights are assigned to the different words using a neural network which are then clustered and analyzed for classification using a clustering algorithm. The final results are averaged for all reviews for a particular hotel to determine its score [4]. Figure 2 shows a Customer Response Evaluation System

Marketing too profits from the availability of such data by using search data generated by aggregators such as Expedia and TripAdvisor to develop discount programs and other offers to attract customers. The data history of customers available with hotels can also be used to analyze the requirements of customers at particular times and seasons of the year to create effective marketing strategies. Loyalty programs can be developed to retain long-term customers which can be identified using this data. Events and important occasions can be kept track of in order to release special offers, promotions and advertisements. Data gathered from social media sites is one which can be used most efficiently, through sentiment analysis techniques, to modify marketing strategies according to the different customer demographics.

3.2 Yield Management

Yield or Revenue Management deals with price optimization of the different resources offered by a hotel according to different internal as well as external factors. These factors could be the weather or season, the demand and supply in the local market or any internal pricing strategy being implemented.

As mentioned earlier, revenue management was among the first areas where business intelligence was used. Traditional revenue management tools were improved considerably with the advent of big data. Data available on booking sites and on search engines provided different customer channels, resulting in more sources of revenue but also more complexity in the economic management of a hotel. This data however could be leveraged to gain insights about the different customer channels and types so as to align the revenue system accordingly. One example of the above is the *innRoad* Real-Time Revenue Management System [3].The innRoad system consists of three components - a forecasting module, a network optimizer and a suite of channel-level optimization modules. It uses real-time data to forecast property demand rates according to different segments and dates and then uses these forecasts for economic evaluation and allocation of rooms. The channel optimization modules use these evaluations along with the data they receive from various channels (rankings, reviews, etc.) to generate real-time prices for different channels, thus providing valuable information regarding the demand and supply view for the hotel [3]. Figure 3 shows the innRoad Real-Time Revenue Management System

The result of such an optimized system is an increase in revenue and decrease in management time and effort. The variety of data being generated online can be gathered and structured to create a big data warehouse which can be leveraged by such revenue systems to provide fast and useful real-time business analytic solutions.

3.3 Operational Effectiveness

There are several internal operations such as energy and water consumption, which can be optimized to ensure effective resource planning in any hospitality industry. Big data can prove to be a major tool in this area too, especially for big chains. Internal data from hotels belonging to a particular chain can be consolidated and analyzed to gain information about the resource utilization by different hotels and in different areas. Necessary strategies can

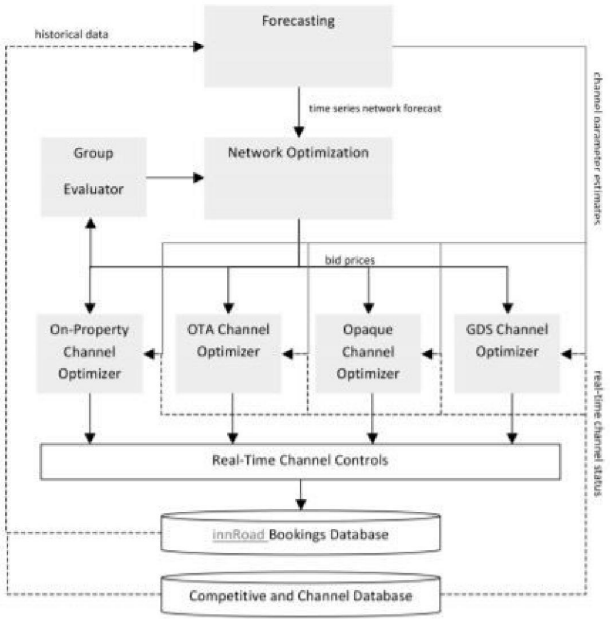

Figure 3: innRoad Real-Time Revenue Management System [3]

then be devised to ensure optimum use of resources.

As mentioned by Kahn and Liu, the administrative data from a hotel can act as "laboratory" for researching different methods to improve energy efficiency [5]. In their study on a major United States hotel chain, they utilized energy consumption data from all hotels in the chain and used a multivariate linear regression technique to understand the factors that affected consumption rate in different areas. One conclusion that they arrived at was that consumption rates were lower in California, which was explained by the stringent energy efficiency laws in California. Conduction of various randomized control-trials also revealed that strategies such as informing people about the downside of increased energy consumption and offering performance bonuses to hotel managers who reduced consumption along with maintaining customer satisfaction, worked in favor of reducing energy costs [5].

The above case study demonstrates how use of even the regular internal energy consumption data generated by hotels over time can prove an essential source of information about the resource utilization in a hotel. Such data mining strategies to contain energy consumption can help decrease not only the operational costs of the hotel but also improve the eco-friendly aspect of the hotel due to reduction of harmful carbon emissions.

4 RECOMMENDATION SYSTEMS

Recommendation systems for travel and hotel sites such as Trivago, Priceline and Expedia are major business intelligence entities in the hospitality sector. These systems deal with huge volumes of data and analyze them to return useful recommendations. The most commonly used recommendation models are - Content-based, Collaborative Filtering and Hybrid. Content-based models provide recommendations based on previous ratings by the user, Collaborative Filtering models provide recommendations based on preferences of similar users and Hybrid models combine the above two i.e. the user's choice as well as the popularity of the recommended hotels [9].

An example of an effective Collaborative Filtering recommendation system is the KASR (Keyword-Aware Service Recommendation) system [9]. The model was implemented by extracting keywords from user comments to generate a list of keywords which were matched with a domain thesaurus (both for the current user and previous users). Similarity between the current user and any previous user recommendations were calculated using approximate and exact similarity methods. Weights were assigned according to the number of such keywords and personalized ratings were calculated to generate recommendations for the current user. The system was implemented on Hadoop and using MapReduce for better scalability [9].

Similar to above, recommendation systems for the other two models, using the original algorithm and modified, have also been implemented. These recommendation systems provide users with the convenience of searching and booking optimum travel and stay packages in one go. Hopper, a useful travel application, uses its predictive analytics to inform users of the best time to fly to get cheaper rates on airline tickets. Doing so, it raised 16 million in a growth funding round in 2016 [10]. Expedia, a well-known travel application, partners with 231,000 hotels and 400 airlines to provide useful deals to customers [13]. These recommendation systems not only serve as useful tools for their customers but also valuable sources of business data for their client i.e. the hospitality sector.

5 CASE STUDIES

Big data analytics has increased tremendously across all service sectors, and the hospitality sector is moving further up the ladder in this era of business intelligence. However, there are a some pioneers which must be mentioned as shining examples in this race. Red Roof Inn, a US economy hotel chain, struck upon the idea of having hotels close to the airport in the winter of 2013-14, as the flight cancellation rates were around 3 per cent at that time and travelers searched for hotels nearby to stay. They used data available publicly regarding flights and weather conditions to launch a marketing strategy that resulted in a 10 per cent increase in business in the targeted areas [7]. Starwood Hotels and Resorts, a large chain with around 1,200 hotels around the world, used local and worldwide market data along with seasonal weather data to update their pricing system and launch marketing campaigns. This resulted in a 5 per cent increase in their revenue-per-room [7]. Marriott Hotels, present at over 3,500 locations and generating 12 billion USD in revenue, revealed their success strategy as being "driven by internet availability". The launch of the *Marriott Reward Program*, based on a business intelligence system which gives real-time information on the member loyalty status, duration of stay, and possible pricing models, has greatly boosted customer satisfaction [6]. InterContinental Hotels in San Fransisco gathered data regarding the energy profiles of their hotel buildings and leveraged it to reduce their

energy costs by 10-15 per cent [14]. Choice Hotels improved their business intelligence program by incorporating *Business Objects*, a business intelligence focused company, in their process to attain real-time data about revenue and occupancy rates. The dashboards with all the key data was provided to all their executives to assist in their decisions, thus empowering the company from within using big data analytics as a tool [6]. These case studies truly reflect the impact and growth of big data in the hospitality sector.

6 CONCLUSIONS

Big Data has revolutionized the field of Travel and Hospitality, and continues to grow as a major factor in all business decisions. What started as a simple revenue management structure, has grown into an intelligent, sustainable system affecting more than one area in the hotel management arena. Launching marketing campaigns, deciding prices and resource allocations, optimizing energy and water consumption, renovating the IT structure, and improving customer comfort and satisfaction, have all been transformed by the arrival of Big Data and business intelligence. There are challenges still to the smooth integration of business intelligence into the day-to-day processes in the Hospitality Industry - overcoming a silos mentality, improving technical expertise to implement the complex Business Intelligence infrastructures needed, and gathering the resources required to support these structures. However, we have observed how the use of big data analytics and intelligence has reformed the hotels who used them. Integration of Big Data and intelligent systems as part of the decision-making process has the potential to become the next 'big thing' for the Hospitality Sector.

REFERENCES

[1] Duetto. 2015. *Bringing Predictive Analytics to the Hotel Industry.* Technical Report. Duetto.

[2] Justin Guinn. 2017. Business Intelligence Tools. (2017).

[3] innRoad. 2015. *Big Data Revenue Management for Independent Hotels.* Technical Report. innRoad.

[4] Ming-Shen Jian, Yi-Chi Fang, Yu-Kai Wang, and Chih Cheng. 2017. Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud, In 2017 19th International Conference on Advanced Communication Technology (ICACT). *International Conference on Advanced Communications Technology(ICACT)*, 791–795. https://doi.org/10.23919/icact.2017.7890201

[5] Matthew E. Kahn and Peng Liu. 2016. Utilizing "Big Data" to Improve the Hotel Sector's Energy Efficiency: Lessons from Recent Economics Research. *Cornell Hospitality Quarterly* 57, 2 (2016), 202–210. https://doi.org/10.1177/1938965515619489

[6] Diane Korte, Thilini Ariyachandra, and Mark Frolick. 2013. Business Intelligence in the Hospitality Industry. *International Journal of Innovation, Management and Technology* 4, 4 (2013), 429–434. https://doi.org/10.7763/IJIMT.2013.V4.435

[7] Bernard Marr. 2016. How Big Data And Analytics Are Changing Hotels And The Hospitality Industry. (2016).

[8] Mauricio. 2016. The role of big data in the travel and hospitality sector. (2016).

[9] Shunmei Meng, Wanchun Dou, Xuyun Zhang, and Jinjun Chen. 2014. KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications. *IEEE Transactions on Parallel and Distributed Systems* 25, 12 (2014), 3221–3231. https://doi.org/10.1109/TPDS.2013.2297117

[10] Sarah Perez. 2016. Hopper raises 16 million for a travel app that tells you the best time to fly. (2016).

[11] Gloria Phillips-Wren and Angela Hoskisson. 2014. Decision Support with Big Data: A Case Study in the Hospitality Industry. *Frontiers in Artificial Intelligence and Applications* 261, DSS 2.0 – Supporting Decision Making with New Technologies (2014), 401–413. https://doi.org/10.3233/978-1-61499-399-5-401

[12] Gil Press. 2014. 12 Big Data Definitions: What's Yours? (2014).

[13] Isabel Thottam. 2017. How Expedia, Hopper and Skyscanner Use Big Data to Find You the Cheapest Airfares. (2017).

[14] Mark van Rijmenam. 2017. Why Hotels Should Apply Big Data Analytics To Provide a Unique Guest Experience. (2017).

Big Data Analytics in Tourism Industry

Weixuan Wang
Indiana University Bloomington
Bloomington, Indiana 47405
wangweix@indiana.edu

ABSTRACT

This study focused on how the tourism industry has been impacted by the development of the Internet and improvements in information and communication technologies. This study explored how big data are generated related the tourism industry and how big data analytic has influenced and can further affect tourism research.

KEYWORDS

i523, HID234, Big data, Tourism industry, Tourism research

1 INTRODUCTION

Information Communication Technologies (ICTs) have been transforming tourism business globally and revolutionizing the world of Tourism. It transforms tourism from a labor-intensive to an information-intensive industry [11]. Tourists influence by the developments in search engines, network speed and capacity have been using use technologies for better planning and experiencing their trips [13]. In addition, ICTs enable travelers to access reliable and accurate information and make reservations faster, cheaper and more convenient than the traditional way [3].

The development of ICTs also enables Internet users to both create and distribute information (especially multimedia information), which is called user-generated content (UGC) or consumer-generated content (CGC) [3]. Platforms for UGC or CGC such as blogs, virtual communities, wikis, social networks, collaborative tagging (bookmarks), and media sharing sites play an increasingly important role as information sources for tourists[12]. Social networks like TripAdvisor, Instagram, Facebook, Yelp, and booking.com are essential for tourists for multiple reasons: for their preparation of trips (booking hotels), during a trip (choose restaurants) and sharing their experience (writing hotel or restaurant reviews). Millions of data records are produced daily in regards to tourism by tourists, businesses and public services [9]. These data can be distinguished as big data by its volume, velocity (the speed it is produced), variety (different formats), variability (diversity of sources) and volatility (different level of production) [8].

Big data has been attracting more and more attention from tourism business practitioners and tourism researchers alike [4]. Big data analytics which is the "activities of the specification, capture, storage, access, and analysis of such data sets to make sense of its content" [8], provide new opportunities and challenges for tourism practitioners and researchers to understand tourists' behavior. This study explores how big data are generated in the tourism industry and used in tourism research, further explore the implication and influence of big data and big data analytics for future research.

2 BIG DATA IN THE TOURISM INDUSTRY

Most activities in the tourism industry had been generating a huge amount of data for several years. Booking flight tickets, reserving a hotel room and renting a car all leaves a data trail [9]. These data could add up to more than hundred of terabytes or petabytes structured data in the conventional databases [1]. Discussions of travel planning on online travel community, status updates and posts on social media like Facebook and Twitter, compliments and compliant on review websites like TripAdvisor constructs more challenging and live unstructured data that arrives at a much faster pace than a conventional database [1]. Tourism practitioners are trying to understand tourists' behavior by accepting and analyzing these big data [9].

Airline and hotel chains have been using their big data which is the large volume of structured information that has been produced internally [8]. Airlines and hotels have been using this tool to analyze prices of plane ticket and hotel room [2]. Moreover, airlines have optimized the details of planning for the crew and routing [2, 9]. The online sector of the tourism industry has also quickly adopted big data to improve internal decisions and understand customers [1]. The online sector of the industry include meta-search engines (like Google), online travel agencies (like Expedia) and some information website companies that distribute tourism information (TripAdvisor)[8]. For example, Amadeus has developed a program "Amadeus Airline Cloud Availability" that can generated special result and increase search for its customers and Kayak has developed a program to predict costs and prices for tourists[9].

2.1 Use of Social Media data in the Tourism Industry

Tourists in the digital age often use a variety of tools to access information that the tourism industry or other users have provided [12]. A tourist produces a high volume of data when they are searching for travel websites, reporting issues on mobile applications, sharing traffic information in the cities, searching and posting on social media, taking and sharing photos, reporting experience on travel websites and social media [1, 9]. All these data that are produced constantly can demonstrate tourists' motivation, interests, and their planning patterns and so on [13].

Previous studies have demonstrated several different usage and formats of big data in the travel and tourism industry [13]. Social media is one of them that has a huge effect on the tourism industry. Social media includes social networks, review sites, blogs, media sharing, and wikis [12]. The exceptional growth of these data sources has inspired companies and institutions to come up with new strategies to understand the socio-economic phenomenon in various fields [9]. Discussions and information sharing on social media are considered as electronic word-of-mouth (eWOM) that has

in some degree substituted tradition face-to-face word-of-mouth (WOM) for information exchange of tourist experience [3]. According to a study on travelers' consulting with social media for travel planning in the US in 2014, 44 percent of people who are within the age group 18-34, use information in social media before planning for travel [10].

Photo post on photographic sharing website also can also provide extensive information on the tourists. Previous studies have connected photos posted on Panoramio, Flickr, and Instagram [2, 8]. Because when a tourist post pictures on these websites, their photo is tagged with geographic locations and ordered chronologically. Therefore analyzing photos posted by tourists can provide a photo density map to better understand tourists' behaviors, and potentially provide opportunities to detect atypical tourists behavior and characterize communities behaviors. However, the study also has its own limitation because of the limitation of technology to better exploit the data [2]. Another study focused on the sequence of locations in shared geotagged photos by tourist to identify and recommend travel routes which helped the travel recommender system to generate personalized recommendation according to interests and time available [6].

2.2 Other Big Data in the Tourism Industry

Beside the use of social media content to analyze tourists behavior, previous studies by Statistics Netherlands has also proposed using other innovative ways to understand tourists behavior by using mobile phone [5]. First method is using log data collected by an app installed on mobile devices, which allowed researchers to tract accurate movements of a person or family [5]. This app also can pop up different questions that be triggered by location or change of time, such as trip purpose, service satisfaction and activities [5]. This innovative design combined the traditional survey with log data from smart phone measurements produced a rich and valuable sets of data [5]. However, this kind of method may be hard to get willing participants, because of privacy concerns and also technical issues such as people may not know how to download and use such application.

Another project from Statistic Netherlands uses aggregated mobile phone meta-data based on call detailed records from 2012 to 2014 [5]. This study collaborated with two telecom providers. Call detailed records contained information of the date and time and location where a communication through mobile network is used. The study uses these information and roaming data to identify unique foreign tourists, was able to detect different groups of foreign tourists and what are their favorite touristic sites within Netherlands [5]. The limitation of this research is also restricted because it requires collaboration with telecom providers and its privacy concerns. With the technology development and widespread of WI-FI, when tourists go to another country they may not need to have roaming service in their destination [5].

3 BIG DATA IN TOURISM RESEARCH

Although tourism scholar has recognized the importance of UGC data such as travel blogs, online reviews and social media post as a form of eWOM has a huge influence in creating destination image [3, 12]. Tourism scholar has also done content analysis on online reviews and travel blogs, but recognizing big data and using big data in tourism research is still limited [3, 11].

Most tourism research utilizing big data are still focusing on CGC or UGC, especially online reviews for a hotel. A recent study conducted by Guo, Barnes and Jia used data mining approach and linguistic analysis to extract meaning from 266,544 online reviews for 25,670 hotels [4]. They mined their customer review data from TripAdvisor using a web crawler [4]. Through their linguistic analysis of their data and cross-comparing with perceptual mapping of the hotels, they found 19 controllable dimensions that are important for hotels to manage their interactions with visitors (such as the price for value, check in and check out) [4].

Another study also focused on UGC and trying to find out determinants of hotel customer satisfaction by dividing customers into different by language group [7]. This study collected 412,784 reviews on TripAdvisor for 10,149 hotels in China. They have found out that tourists speaking different languages (such as Chinese, English, German, French, Russian etc.) differs significantly in terms of their emphasis on various attributes of hotels, and forming different satisfaction rating for hotels [7].

Both of the two studies mentioned above were from tourism or hospitality journals, were conducted by tourism researchers. Another study from outside of tourism research cohort provided a different study using big data to understand tourist behavior. This study designed and evaluated a big data analytics method using geotagged photos shared by tourists on Flickr to support destination management organization in analyzing and predict tourist behavior patterns at destinations (for this study itis Melbourne, Australia). The study designed a geotagged photo analytic artifact with textual meta-data processing geographical data clustering, representative photo identification and time series data modeling. This study demonstrated how to analyze unstructured big data to enhance strategic decision making in tourism destinations, provided insight on how city tour can be designed to better reflect tourists' interests and enrich their travel experience [8].

4 CONCLUSION

This study has explored the literature of big data and its implication in the tourism industry. Both tourism practitioners and tourism researcher has recognized the influence of big data and big data sources for tourism development. Big data in the tourism industry are generated by tourists directly, compared to traditional data sets that are gathered from surveys. Therefore, big data presented us opportunities to better understand tourist behavior, their motivations, and interests. However, big data also poses challenges for tourism practitioner and tourism researchers.

Like these two studies from tourism and hospitality journals, they share similarities in terms of data collection methods. Tourism researchers have recognized the importance of user-generated data which was able to provide them the volume of data they need for better generalization. One limitation of this kind of tourism research is that they only focus on hotel reviews, but their method could extend to other tourism sectors such as attraction and event to evaluate or review the dimensions of tourist satisfaction. Another limitation they have is that they are only focusing on the text-based data from review website. How to integrating and getting useful

information from other unstructured data such as image, video, post on Facebook and Twitter is still challenging for tourism researchers. However, studies outside of tourism domains can be helpful in helping tourism researchers to utilize other formats of big data to understand tourist behavior. Therefore, collaboration with other fields and utilizing unstructured big data, and big data analytics in relation to tourism are much needed for tourism research.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Rajendra Akerkar. 2012. *Big Data & Tourism.* Technical Report. Technomathmatics Research Foundation.

[2] G. Chareyron, J. Da-Rugna, and T. Raimbault. 2014. Big data: A new challenge for tourism. In *2014 IEEE International Conference on Big Data (Big Data).* IEEE, Washington, DC, USA, 5–7. https://doi.org/10.1109/BigData.2014.7004475

[3] Jin Chung and Dimitrios Buhalis. 2009. *Virtual travel community: bridging travellers and locals.* IGI Global, USA. 130–144 pages. http://www.ebook.de/de/product/8468470/tourism_informatics_visual_travel_recommender_systems_social_communities_and_user_interface_design.html

[4] Yue Guo, Stuart J. Barnes, and Qiong Jia. 2017. Mining meaning from online ratings and reviews: Tourist satisfactionanalysis using latent dirichletallocation. *Tourism Management* 59, Supplement C (2017), 467 – 483. https://doi.org/10.1016/j.tourman.2016.09.009

[5] Nico Heerschap, Shirley Ortega, Alex Priem, and May Offermans. 2014. *Innovation of tourism statistics through the use of new big data sources.* Technical Report. Statistics Netherlands.

[6] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. 2013. Travel route recommendation using geotagged photos. *Knowledge and information systems* 37, 1 (2013), 37–60.

[7] Yong Liu, Thorsten Teichert, Matti Rossi, Hongxiu Li, and Feng Hu. 2017. Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tourism Management* 59, Supplement C (2017), 554 – 563. https://doi.org/10.1016/j.tourman.2016.08.012

[8] Shah Jahan Miah, Huy Quan Vu, John Gammack, and Michael McGrath. 2017. A Big Data Analytics Method for Tourist Behaviour Analysis. *Information & Management* 54, 6 (2017), 771 – 785. https://doi.org/10.1016/j.im.2016.11.011 Smart Tourism: Traveler, Business, and Organizational Perspectives.

[9] S. Shafiee and A. R. Ghatari. 2016. Big data in tourism industry. In *2016 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC).* IEEE, Isfahan, Iran, 1–7. https://doi.org/10.1109/ECDC.2016.7492979

[10] Statistica. 2014. Travelers who consult social media when travel planning in the United States as of April 2014, by age group. (2014). https://www.statista.com/statistics/305150/travelers-using-social-media-for-travel-planning-by-age-us/ accessed 2017.

[11] N.L. Williams, A. Inversini, N. Ferdinand, and D. Buhalis. 2017. Destination eWOM: A macro and meso network approach? *Annals of Tourism Research* 64 (2017), 87–101. https://doi.org/10.1016/j.annals.2017.02.007 cited By 0.

[12] Zheng Xiang, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* 44, Supplement C (2015), 120 – 130. https://doi.org/10.1016/j.ijhm.2014.10.013

[13] Karen L. Xie, Kevin Kam Fung So, and Wei Wang. 2017. Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management* 62, Supplement C (2017), 101 – 110. https://doi.org/10.1016/j.ijhm.2016.12.004

Big Data in Recommendation System

Yujie Wu
Indiana University Bloomington
Bloomington, Indiana 47401
yujiwu@iu.edu

ABSTRACT

This paper will befocused onhow the company recommend their products and services to their customers based on the dataabout customer's preferences through the case of Netflix and Yahoo.

KEYWORDS

Recommendation system, Netflix, Yahoo

1 INTRODUCTION

With the development of web technology and online business, recommender systems are widely used in e-Commercial business platform such as Amazon, eBay, Monster and Netflix. They process a tremendous number of online commercial activities and provide the personalized online business experience, which relies on the recommender system. The recommender system tells the customers what they are looking for, what they want to buy, and so on. With the recommender system, less popular products can attract peoplefis attention and e-Commercial business model works more efficient and profitable.

The basic problem of recommendation system is personalized matching of items (people, products, services, jobs, etc.) to people[3]. The recommendation system takes the data that produced by peoplefis online activities and some specific criteria such as overall context, user information, community information, properties of items plus the machine learning or data mining algorithms to output some information or suggestion that people may be interested in or related to according to their preferences for some goals[3].

2 NETFLIX RECOMMENDER SYSTEM

Netflix recommender system is an industrial-scale and real-world recommender system. Its recommendation is based on personalization. For customers (household), everything that they could see on the front-end webpage containing rows and columns is recommendation. The columns are sorted by the ranking and diversity. The personalized genre rows focus on user interest which is important for user satisfaction. They are generated based on the usersfi recent activities, ratings, comments, or usersfi preference settings. The recommender system will filter out the movies if they have been watched before and exclude duplicated tags and genres when providing recommendations and suggestions[3].

Netflix recommender system just takes the user preferences and output the immediate recommendations. How is it highly related to Big Data? According to the statistics from Netflix, last two quarters in 2013 have four million new registered subscribers, which leads to total 29.2 million subscribers were actively using Netflix. There are 4 million ratings, 3 million searches, 30 million plays happening in Netflix website every day. At the end of 2013, Netflix reached 44 million members[3]. A large amount of data is collected each day. Therefore, it becomes reality that Netflix recommender system could use Big data which usually beats better algorithms to provide recommendations.

The algorithms that Netflix recommender system is using are Restricted Boltzmann Machines (RBM) and a form of Matrix Factorization. They are developed as part of the Netflix 2007 Progress Prize which worth several million dollars. Restricted Boltzmann Machine is a neural network. The form of Matrix Factorization is an asymmetric form of SVD which can take implicit information into account[1]. Both algorithms consist of a tremendous number of different machine learning techniques. The algorithms consume a large amount of data as their input. Machine learning techniques form an abstract model which is waiting for data stream to shape it. Once the model reaches convergence or it becomes mature enough, the algorithms output the prediction which is used as the recommendation for Netflix users. More data means more precise the outcome is.

The recommendation algorithms are designed based on the hypothesis that the suggestions will increase the member engagement with Netflix service and ultimately attract more users and more profits. To verify whether the algorithm works as expected, Netflix designed a test, named AB test. AB test is an experimental approach to figure out the changes of webpages which maximize an outcome of interest. The test contains two identical versions with only one different variation which possibly affect customer's behavior[3]. For instance, the A version of a website has some webpages that could be accessed through a category list. The version B of that website is modified from version A that the webpages which can be accessed only through a category list now have their own shortcuts listed on the main page of the website. Once executing the AB test, it is obvious whether the modification on that variation increases the user engagement.

To modify the webpage, it should measure or evaluate all related metrics, which is a data-driven process. Metrics could

be short-term or long-term. Sometimes, short-term metrics do not fit the long-term goals. For example, larger quantity of clicks does not necessary mean better recommendation. However, long-term metrics such as member retention works better in Netflix[3]. With the choice of metric, Netflix monitors how users interact with different algorithms during the testing.

3 YAHOO RECOMMENDER SYSTEM

The main page of Yahoo contains many modules such as advertising module, search queries recommendation, breaking news recommendation, and application recommendation. All recommendations rely on Yahoo recommender system based on the given context such as user data and user preferences. Yahoo recommender system is not merely an algorithm or a piece of code, it is an environment involves items, context, and metric. Items could be articles, advertisements, movies, songs that users may be interested in. Context could be query keywords, pages, mobile, social media that users provided while surfing online. Metric could be click rate, revenue, engagement that needs to be optimized for achieving some long-term business objectives[4].

Every second, a tremendous amount of data from users and machines is feed to the system. It is a problem that big data matters. Therefore, big data analytics and machine learning algorithms can be applied to improve or optimize the metric and the system while recommendation is on-going.

The data is easy to obtain but its quality is not guaranteed since the nature of data resource. Various factors including the properties of the item, context, feedback, and constraints specifying legitimate matches may affect data quality and eventually the solution. Yahoo recommender system uses collaborative filtering to deal with such problem.

Collaborative filtering assigns each item an individual rating to form a consensus recommendation. To be more specific, collaborative filtering has three branches which are user-based collaborative filtering, item-based collaborative filtering, content-based collaborative filtering. As the name implies, user-based collaborative filtering groups the similar users and find their preferences, then it predicts the interest of current user based on the group of the similar users. Item-based collaborative filtering recommends items to current user based on the rating that is assigned to each individual item. Content based collaborative filtering finds the items with the similar properties that the current user likes[4].

Collaborative filtering is now the most prominent approach to generate recommendations. It presumes that the ratings of the items are given by users. Then it takes a table of data including the users and item ratings to compare the values and return the top-ranked items for the current user[4]. Finally, collaborative filtering outputs a prediction that describes how much the current user likes or dislikes the item.

As mentioned before, the input is a table which has a set of attributes. Each attribute represents an item and each tuple represents a user. Therefore, the value in each cell means the rating of the item given by corresponding user. Collaborative filtering finds some most similar users and their items to the current user, then remove the items that current user have already seen or purchased. Hence, the input data table only includes similar users and items which will be recommended to the current user.

Here remains a problem that how to define the similarity between users. Let A and B be two different users and let I be the set of items that both user A and B rated. Let $r_{a,i}$ be the rating of user A for i^{th} item. Let $\overline{r}_a$ and $\overline{r}_b$ be the average value of all items in set I rated by user A. Therefore, the similarity could be calculated as the following function[4]:

$$sim(a,b) = \frac{\sum_{i \in I}(r_{a,i} - \overline{r}_a)(r_{b,i} - \overline{r}_b)}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r}_a)^2}\sqrt{\sum_{i \in I}(r_{b,i} - \overline{r}_b)^2}}$$

The similarity function is called cosine similarity. The function assumes each tuple in the table is a vector. Since similarity function uses cosine value, the possible value of similarity is between -1 and 1. If two vector points to the same direction, cosine similarity value equals 1. If two vector points to the opposite direction, cosine similarity value equals -1.

Once the data table is feed to the algorithm, the prediction of the rating value of some random item i which will be recommended to the current user could be calculated as follows[4]:

$$pred(a,i) = \overline{r}_a + \frac{\sum_{b \in N} sim(a,b)(r_{b,i} - \overline{r}_b)}{\sum_{b \in N} sim(a,b)}$$

where a is the current user and b is a random user in the data table. The set N is group of all users in the data table except the current user. The item with highest rating value will be returned by the algorithm as the ultimate suggestions to the current user.

Yahoo recommender system in advertisement module employs machine learning technologies such as singular value decomposition (SVD) and latent semantic indexing (LSI) to provide recommended keywords. SVD and LSI are also used to recommend music and movies[2]. Like the most machine learning algorithms, SVD and LSI train the model based on the numeric data. Since a tremendous amount of data is gathered in a short period of time, the training time will increment exponentially and leads to a delay in response finally. The solution of Yahoo recommender system is partition the data set and develop new method for certain sets. The training time , as a result, increases log-linearly in practical situations[2].

4 CONCLUSION

Big data is highly involved in recommendation machines. Both Netflix and Yahoo utilize machine learning algorithms such as Restricted Boltzmann Machines, a form of Matrix Factorization, singular value decomposition, and latent semantic indexing. Yahoo also uses collaborative filtering algorithm for item recommendation. Netflix uses AB testing for validating a new recommender algorithm. In the future, more efficient and more elegant algorithms will be invented. Big data will lead to a more precise recommendation.

REFERENCES

[1] Xavier Amatriain. 2014. How does the Netflix movie recommendation algorithm work? Online. (12 2014). https://www.quora.com/How-does-the-Netflix-movie-recommendation-algorithm-work

[2] Dennis Decoste, David Gleich, Tejaswi Kasturi, Sathiya Keerthi, Omid Madani, Seung-Taek Park, David M. Pennock, Corey Porter, Sumit Sanghai, Farial Shahnaz, and Leonid Zhukov. 2005. Recommender Systems Research at Yahoo! Research Labs. Online. (1 2005). https://www.cs.purdue.edu/homes/dgleich/publications/decoste2005%20-%20yahoo%20recommender%20systems.pdf

[3] Geoffrey Fox. 2017. Big Data Applications and Analytics Case Study: e-Commerce and Life Style Infomatics: Recommender Systems I. Online. (9 2017). https://drive.google.com/file/d/0B6wqDMIyK2P7YkIwczVfQlJqVG8/view

[4] Geoffrey Fox. 2017. Big Data Applications and Analytics Case Study: e-Commerce and Life Style Infomatics: Recommender Systems II. Online. (9 2017). https://drive.google.com/file/d/0B6wqDMIyK2P7UVloVElaZ2FXcTg/view

Big Data Analytics in Finance Industry

Gagan Arora
Indiana University
gkarora@iu.edu

ABSTRACT

Discusses the importance of Big data, data classification, cross industry comparison and analytic and its impact on finance industry and how customer satisfaction can be improved to achieve competitive advantage. We will also discuss the qualitative and quantitative aspect of big data in Finance industry and how we can leverage big data in achieving high quality financial products.

KEYWORDS

Big data,finance, i523, data classification,HID301, data type, big data impact, cross-industry comparison, banking, customer personalization, pattern recognition, multi document summarization, structured and unstructured data

1 INTRODUCTION: DISCUSS IMPORTANCE, HOW DOMINATING

By its nature of the business, the finance industry is always driven and dominated by data. The existence of Big data in the finance industry has exposed the big opportunity of growth and value extraction but at the same time imposed the various new challenges, which demand new skill set. [3] suggests that finance experts believe there is a huge potential in terms of value extraction from the financial big data. They also believe that finance industry can benefit more than any other industry. Historically, data was always there in some format either non-digital or digital. However, with digitalization, this data has fallen into the prevalence of high volume of information, which we call as Big Data. Dominant drivers for the actuality of big data in the finance industry are mainly customer call logs, social media, news feed, regulatory data etc. Call logs, news feed and etc. fall into the category of unstructured data which is identified as an area where we can extract vast amount of business value. We will discuss the various types of data, which is being generated at a phenomenal rate and what business value can be extracted out of big data. We will also discuss what all challenges this big data imposes on the finance industry. [4] talks about the effectiveness of extracting value out of big data in the finance industry and utilizing it to improve business operation.

2 MARKET IMPACT: PACE AT WHICH MARKET IS ADDING DATA

[1] talks about the three V of big data in finance industry: volume, velocity and variety. We will also discuss the fourth V aspect of it in a later section, which is a vulnerability. [3] clearly depicts the amount of financial data pouring in the daily basis. TechNaviofis forecast (Technavio 2016) predicts data will grow at a CAGR [compound annual growth rate] of 61 percent over the period of 2017-2021. According to the IDC financial insight 2016, every second there is around 10,000-payment card transaction and this number is expected to double by the end of this decade. The Capgemini/RBS Global payments study for 2012 suggests there was about 260 billion transactions in 2012 and is expected to grow between 15 and 22 percent for developing countries. Main drivers contributing to the big data in the finance industry are Data growth, increasing scrutiny from regulators, digitalization of financial products, changing the business model and increased customer insight platforms such as customer service. [1] shows 76 percent of banks say the business driver for embracing big data is to enhance customer engagement, retention, and loyalty and 71 percent of banks say that to increase their revenue, they need to better understand customers and big data will help them to do so.

3 COMPETITION/PROBLEM: HOW PEER INDUSTRY ADDING DATA AND THEIR INITIATIVE

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

4 WHY BIG DATA IN FINANCE AND ITS IMPACT ON THE CUSTOMER PERSONALIZATION AND THEIR SATISFACTION

Big data technologies not only help in extracting the effective business value but analysis of unstructured data in conjunction with a wide variety of data set also helps in extracting commercial value. Big data in finance industry does not necessarily decode to valuable or actionable information. The real benefit lies in developing the technologies, which can be used to extract business and commercial value. [7] talks about what all advantage we can extract from the big data in the finance industry. Few examples are: Detection of

false rumors that try to manipulate the finance market, Assessment of exposure to a reputational risk connected to consulting service offered by banks to their customer and Discover topic trends, detect events, or support the portfolio optimization or asset allocation. Big data based pattern recognition can also help in enhanced fraud detection systems and prevention capability systems. Other benefits of utilizing big data include building a machine learning based algorithm to achieve higher performance and accuracy in the trading algorithm and Enhanced market trading analysis. There has been proven research [6] which states more data increases accuracy and precision of simulations which is the backbone of financial modeling based analytics. This research [6] states Modern modeling techniques are data hungry

5 BIG DATA CLASSIFICATION IN FINANCE INDUSTRY: STRUCTURED UN-STRUCTURED AND SEMI-STRUCTURED

Financial service system has a varied variety of data pools that are held by various stakeholders. At a high level of abstraction, we can classify them into three major categories: Structured data, Unstructured data, and Semi-structured data. With the emergence of too much data supply, there has been operational intelligence initiative such as firms like SPLUNK which uses data mining approaches to fetch valuable information out of any type of log. There have been studies [2] which shows utilizing structured data for analyzing event logs. Another advantage of structured data is that it makes the concept of Data Virtualization easy as data can easily be virtualized if we have structured data, which in turn make easy to extract patterns. Extracting patterns from the customer banking activities gives banks competitive advantage as they can make better personalized financial products for customers.

5.1 Structured data

This reflects the data which has a higher degree of an organization such as a relational database where information/data is easily searchable and we can easily apply standard algorithm to extract patterns out of it. Examples of such data set include Trading applications, Enterprise finance resource planner, Retail banking systems, Credit history database systems and other financial applications that use legacy application systems. Structured data always has a big advantage of being easily entered, stored, queried and analyzed. Most of the personal banking financial statements are stored in a structured way. Structured dataset combined with the distributed systems can be leveraged to achieve Structured big data set on which we can run optimized SQL queries to retrieve patterns. [5] discusses various SQL based ways to specify information quality in data which can be used to filter out the noise.

5.2 Unstructured data

With the emergence of social media, blogging and mobile usage there has been a phenomenal amount of data which we can classify as unstructured data. Example of unstructured data includes Daily stock feeds, Company announcements, Finance news, Articles, Blogs, Customer feedback/reviews and etc. There have been researches such as multi-document summarization and machine learning algorithms to utilize the unstructured data to extract value. There is a big advantage with the unstructured data that it is a platform, programing language, technology compatible ie. Two or more machines which different platform can interact with each other using unstructured data. This means financial big unstructured data can be stored on distributed systems and pattern recognition application can be used to extract value. There are also other ways such as transforming unstructured data to structured and then fetching intelligence out of it since structured data is akin to machine language.

5.3 Semi-structured data

As the name suggests this data type includes the aspect of both structured as well as an unstructured data type. Examples of semi-structured data includes: Financial products markup language[FpML], Financial Information eXchange[FIX], Interactive Financial eXchange(IFX), Open Financial eXchange(FEDI) , Market data definition language (MDDL) and etc. [4] suggests nowadays semi-structured data dominates the data in finance industry which contributes to around 80-85 percent of finance data

6 VARIOUS CHALLENGES UTILIZING BIG DATA VALUE IN FINANCE INDUSTRY

There are multiple challenges and constraints in extracting value out of big financial data. The biggest challenge is old IT culture and infrastructure. The Much financial organization still uses old IT infrastructure which is not compatible with the big data application thus fail to take advantage of big data. Other challenges include lack of skill set and data privacy and security. With the emergence of digitalization, customer data is saved persistently because of which there has been continued concern regarding the customer privacy. Regulatory bodies guidelines on customer data are always ill-defined because of which is there is always a concern regarding the use of customer data.

7 REQUIREMENTS SOLUTION AND CONCLUSION

In this section, we will discuss various technical requirements needed to achieve value extraction from the big data in the finance industry. There are various technical requirements such as Data Acquisition, Data Quality, Data Extraction, Data Integration, Decision support. In order to fulfill requirements, a hybrid approach combining computer science, algorithms, statistics, data mining, machine learning and pattern recognition study needs to be adopted. To explore the advantage of big data there have been initiatives like data virtualization, multi-document summarization, pattern recognition from LOGS and many start-ups have been emerged. All big companies such as Microsoft, Google, IBM and Amazon are investing heavily in this field to leverage business and commercial value out of it. There has been changed in the industry pattern where financial industry is resorting big data to strategize their business. According to [3] with a very rapid pace, the financial industry is utilizing big data advantage in investment analysis, econometrics, risk assessment, fraud detection, trading, customer interaction analysis and behavior modeling. If we look at the Big promise the Big

data holds in the finance industry, progress in this field is still in nascent stage and we expect more growth in upcoming years.

REFERENCES

[1] Daniel D. Gutierrez. 2014. *Big Data for Finance.* Technical Report. Dell & Intel. https://whitepapers.em360tech.com/wp-content/files_mf/1427803213insideBIGDATAGuidetoBigDataforFinance.pdf

[2] M. Hinkka, T. Lehto, and K. Heljanko. 2016. Assessing Big Data SQL Frameworks for Analyzing Event Logs. In *2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP).* IEEE, http://ieeexplore.ieee.org/document/7445319/, 101–108. https://doi.org/10.1109/PDP.2016.26

[3] Kazim Hussain and Elsa Prieto. 2015. *Big Data in Finance.* chapman and chapman and hall/crc, https://www.cs.helsinki.fi/u/jilu/paper/bigdataapplication04.pdf, Chapter 17, 329–356.

[4] Kazim Hussain and Elsa Prieto. 2016. *Big Data in the Finance and Insurance Sectors.* Springer, Cham, "https://link.springer.com/content/pdf/10.1007/", Chapter 12, 209–223.

[5] A. Parssian, W. Yeoh, and M. S. Ee. 2015. Quality-Based SQL: Specifying Information Quality in Relational Database Queries. *Computer* 48, 9 (Sept 2015), 69–74. https://doi.org/10.1109/MC.2015.264

[6] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. 2014. Modern modelling techniques are data hungry a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 14, 1 (22 Dec 2014), 137. https://doi.org/10.1186/1471-2288-14-137

[7] Sonja Zillner, Tilman Becker, and Munn. 2016. *Big Data-Driven Innovation in Industrial Sectors.* Springer International Publishing, Cham, Chapter 4, 169–178. https://doi.org/10.1007/978-3-319-21569-3_9

Big Data Application in Restaurant Industry

Sushant Athaley
Indiana University
sathaley@iu.edu

ABSTRACT

Big data application is not only getting used in scientific research but it is also getting used commercially. Most of the businesses are using big data to change the way they are operating and getting rewarded. The restaurant business is also currently evaluating how big data can be used. This study focuses on the big data elements for the restaurant industry, gathering of big data, analytics, available big data solutions, current implementations, and challenges faced by restaurant industry in big data application. This study considers information from various sources like articles, books and web to provide this information.

KEYWORDS

i523, hid302, big data, restaurant, application, analytics

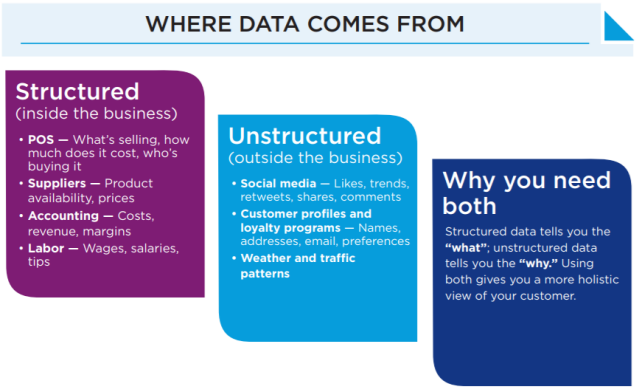

Figure 1: Restaurant - Data Sources [8]

1 INTRODUCTION

Big data is revolutionizing the way business is getting conducted in various industries. The retailer like Amazon uses it to provide personalized buying suggestions and social networking site like LinkedIn uses it to connect more people. Question is, do we have big data available for the restaurant industry and how big data application is going to be beneficial? The restaurant industry is facing challenges like shrinking labor pool, moderate economic growth, costly labor, challenging profit margin, high competition, moderate sales growth and growing expectation from the customer on the dining experience, can big data application help overcome these challenges? [9]

The study is structured as follows. Section *Ingredients* captures various data points available in the restaurant industry for the big data analysis. Section *Consume* provides details on how data can be gathered in the restaurant industry. Section *Recipe for Success* captures various big data analytics which can help to solve different problems. Section *Kitchen Tools and Gadgets* provides information on current big data solutions and tools available for the restaurant industry. Section *Flavourful Implementations* provides real-life examples of big data applications in the restaurant industry. Section *Hell's Kitchen* capture various challenges involved in using big data for the restaurant industry. Finally, section *Conclusion* concludes the study.

2 INGREDIENTS

To understand how big data analytics will help, we first need to find out what are the data points present in the restaurant industry which can be considered as big data. As one of the V-variety of big data, the restaurant also has structured and unstructured data. Structured data is something which is getting generated inside the restaurant and unstructured data is something which is outside of the restaurant. Figure 1 shows restaurant industry data sources.

2.1 Structured Data

Structured data is well formatted, easy to understand and analyze. Restaurant POS (Point of Sale) system shows what's selling, where, and at what time [7]. Food and beverage cost, labor cost, product mix, rent cost are obvious data points. Raw material required for preparation, menu, ingredient consideration, meal preparation, product availability from the supplier, prices of products are the data points which comes from the kitchen of the restaurant. Staffing schedule, table turnover, bar management, wages, salaries, tips, customer feedback is valuable data. The number of time employee coming late, number of times drinks provided as comp due to server error is data [8].

2.2 Unstructured Data

Unstructured data is un-formatted, difficult to gather and analyze. Data shared from social media like trends, retweets, shares, and comments categorize as unstructured data. Customer promotions, customer profile like age, gender, address, email, taste preference, favorite dish, various milestones like birthdate, anniversary, along with family information is also an unstructured data. Weather and traffic information also constitutes as an important data to consider [8].

3 CONSUME

These various data attributes can be collected from the different systems. Most of the data is generated inside the restaurant by the system like POS which captures all sales transactions. POS system can also break down sales by time, size of the party, menu items, and ingredients. The inventory provides information on suppliers, food, beverages, and gas and electricity bill. Payroll provides information on wages, salaries, employee schedule, and time off by the employees. Loyalty program and marketing promotions provides data regarding marketing of the restaurant.

Outside data can be gathered through the various applications like OpenTable, Facebook, Twitter, Yelp, TripAdvisor, Foursquare, Urbanspoon or Instagram, weather and traffic sites. Information can be gathered from customer like his favorite menu/drink item, favorite table, special request, allergies, liking to the presentation, feedback on ambiance, service and food [8].

4 RECIPE FOR SUCCESS

Benjamin Stanley, co-founder of Food Genius, suggests "A restaurant operator shouldn't just jump into big data unless they have a problem they are trying to solve" [5]. Big data analytics can help with various analysis which can solve different issues but it's important to know the problem which needs to be solved. If the goal is to reduce costs, streamline operations or better manage the staff then analysis needs to be done in inventory control, supply chain management or scheduling solutions. If the problem is related to food trends, menu options and improving the customer experience, then focus probably can be on social media, customer demographics, and dining-out trends [8].

Menu analysis can help with deciding the cost of the item, popular menu item, how often items are ordered, the time when menu item ordered, ingredient used and if any ingredient needs to be substituted [5].

Labor cost can be managed better by analyzing overtime pay, absenteeism, costs to sales, costs by department and server, tips, amount of time spent at the table, types of entrees sold and whether the server sells the special. This analysis can be used to motivate, train and provide incentives to the servers [8],[5].

Guest check analytics can help determine what sells well, how often somebody orders certain items and detailed pricing analyses [8]. Customer profile analysis gives insight on demographics of the customer, ages, income level, their family information, kind of food they like, allergies, drink habits, places they dine out, special occasions and this analysis can be used to provide the personalized experience to the customer [8]. Servers can use customer profile analysis to suggest menu choices, celebrate birthdays or special occasions, or run specials to drive more business. Reservation system data analysis helps in understanding who all are coming, when they last visited, what they tend to order, are they celebrating any special occasion and accordingly then chef can decide on the menu [11].

Data mining of data from social media like Facebook, Twitter, Instagram, YouTube can help in understanding sentiments of the customer, social news, trending topic, views on self and competitor restaurants, identify brand or restaurant fans [4]. This mining also provides the capability to get feedback real time and respond at the same time. This information can be used to do targeted marketing for the specific audience [4].

5 KITCHEN TOOLS AND GADGETS

Fishbowl provides cost-effective data analytics solution to the restaurant industry using Hadoop and other technologies. Fishbowl integrated Hadoop with their marketing platform to provide guest analytics, menu management, media analytics, promotions and mobile platform to provide complete solutions [3][2].

MyCheck and MarketingVitals.com together provide mobility and data analytics platform for the hospitality industry [10].

Dickeys Barbecue Pit restaurant has worked with big data and business intelligence service provider iOLAP to develop a proprietary system called as Smoke Stack. Smoke Stack provides real time data analytics to take better decisions [6].

Upserve, a restaurant management platform, provides payment processing, point of sale, data insights to boost margins and exceed guest expectations [11][12].

Founding Farmers gathers data together from Swipely, OpenTable and analytics service Avero Slingshot to do the customer profile analytics and builds top 100 customers to reach out to them in a highly personalized way. It also helps in understanding customers food and drink preferences along with how they would like to be served so that their dining experience can be personalized [5].

6 FLAVORFUL IMPLEMENTATIONS

"A quickservice chain monitors its drive-thru lanes to determine which items to display on its digital menu board. When lines are longer, the menu features items that can be prepared quickly. When lines are shorter, the menu features higher-margin items that take a bit longer to prepare. Those subtle changes in the menu board wouldn't be possible if the company couldn't tap into a steady stream of data in real time to make instantaneous adjustments" [8].

"Haute Dogs and Fries, a two-unit, quickservice restaurant in Alexandria, Va., leverages social media to connect with customers. Being small and community-focused allows the operation to quickly identify market trends and make offers in real-time, says co-owner Lionel Holmes. He monitors social media throughout the day and might post a lunch special at 11 a.m. or a dinner offer at 3 p.m. based on what is trending. Haute Dogs and Fries is on Twitter, Facebook and Instagram and uses email to reach customers and build loyalty" [8].

"Fig and Olive, a seven-location New York-based restaurant group, has used guest-management software to track more than 500,000 guests and $17.5 million in checks. The restaurants have been able to customize the dining experience for individual guests and deliver results with targeted email communications. It's *we miss you campaign* offered complimentary crostini to guests who hadn't dined there in 30 days. The result: Almost 300 visits and more than $36,000 in sales, translating into a return of more than seven times the cost of the program. Matthew Joseph, who leads technology and information systems for the company, says linking POS data with online reservations, plus monitoring social media mentions on Facebook, Twitter or TripAdvisor, helped Fig and Olive create its brand identity and build loyalty" [8].

Dickeys Barbecue Pit, which operates 514 restaurants across the U.S., uses Smoke Stack system to provide near real-time feedback on sales and other key performance indicators. All of the data is examined every 20 minutes to enable immediate decisions. If the sale is not at certain baseline at a certain store in the region then it enables them to deploy training or operation directly to that store. For example, if there is lower than expected sales one lunchtime, and have an amount of ribs there, then text invitation is sent to people in the local area for ribs special to both equalize the inventory and catch up on sales [6].

"Andy Husbands, chef-owner of Tremont 647 restaurant, uses a management system called Upserve to keep tabs on what his customers like and don't like. The software pulls together streams of information like transaction data, OpenTable reservations, and sales history and displays everything on a dashboard that Husbands can access on his phone, giving him insight into how his food and staff are performing. He can instantly see which server has the highest check average and whether it's because server, for instance, sells more appetizers or drinks than his/her co-workers" [11].

7 HELL'S KITCHEN

The restaurant industry is very slow in terms of adopting or spending on new technologies due to small profit margins, high employee turnover and the overall cost of implementation [11]. Most of the restaurants are still using legacy software packages which are inadequate in dealing with the big data. These legacy software packages are cumbersome to upgrade or integrate with new technologies or data streams which are required for the big data analytics. It can take a lot of times to get data from old restaurant software to the data warehouse. Even if data is centralized, it's difficult for most of the restaurants to hire a data scientist to analyze data due to their costly salaries. Only big restaurant chain can afford such costly labor and tools needed for the big data application [1]. Another major challenge is the variety of big data source and format involved in restaurant industry like structured data in form of POS, inventory systems and unstructured data like social networking site or weather reports. Combining data from such various sources is big deal. There are financial challenges also as technology offered to work with big data is expensive which makes leveraging big data challenging for most of the restaurants [3]. Dealing with customer personal data poses a security risk. This sensitive information if collected need to be protected so that it is not misused for identity theft or some other fraud [8].

8 CONCLUSIONS

Big data application offers ample opportunities to solve the various problems faced by the restaurant industry. It is opening avenues which cannot be imagined earlier but adoption of big data application is a bit slow in restaurant industry compared to other industries like retail due to low-profit margins and high application cost. Currently, big data is mostly used by the large chain and Michelin star restaurants who can afford the big data solutions. Efforts are getting made to provide low-cost solutions so that small and medium restaurant can also embrace the big data. There is no doubt that big data application is going to change the way people dine out and as quickly restaurant adopts it the quicker it's going to provide customers that Umami effect.

ACKNOWLEDGMENTS

The author would like to thank internet fraternity who generously contributes information on the web for others enlightenment. The author would also like to thank Dr. Gregor von Laszewski for his review and suggestions.

[3]

REFERENCES

[1] Dipock Das. 2015. Big Data's Last Crusade: Restaurants still slow to embrace smart technology. (May 2015). https://www.hotschedules.com/news/big-datas-last-crusade-restaurants-still-slow-to-embrace-smart-technology/
[2] Fishbowl. 2000. Fishbowl. (2000). https://www.fishbowl.com
[3] Dev Ganesan. 2015. How Big Data Technologies Are Revolutionizing Restaurant Marketing. (Feb 2015). https://www.foodnewsfeed.com/fsr/vendor-bylines/how-big-data-technologies-are-revolutionizing-restaurant-marketing
[4] Lisa Jennings. 2015. Making big data small. *Nation's Restaurant News* 49, 7 (May 2015), 22–23.
[5] Amanda C. Kooser. 2013. BIG DATA. *Restaurant Business* 112, 9 (September 2013), 24–31.
[6] Bernard Marr. 2015. Big Data At Dickey's Barbecue Pit: How Analytics Drives Restaurant Performance. (Jun 2015). https://www.forbes.com/sites/bernardmarr/2015/06/02/big-data-at-dickeys-barbecue-pit-how-analytics-drives-restaurant-performance/ Forbes Article.
[7] John Morell. 2013. Get a Grip on Big Data. (may 2013). https://www.qsrmagazine.com/operations/get-grip-big-data
[8] National Restaurant Association. 2014. Big Data and Restaurants: Something to Chew On. Web. (11 2014). https://www.restaurant.org/Downloads/PDFs/BigData
[9] Restaurant Org. 2016. Restaurant industry to navigate continued challenges in 2016. (02 2016). http://www.restaurant.org/News-Research/News/Restaurant-industry-to-navigate-continued-challeng
[10] Taylor Szabo. 2015. MyCheck and Marketing Vitals Announce Integration of Big Data Restaurant Analytics and Mobile Payment Technology Platforms. Horn Group for MyCheck. (Sept 2015). http://www.businesswire.com/news/home/20150916005807/en/MyCheck-Marketing-Vitals-Announce-Integration-Big-Data
[11] Nicole Torres. 2016. How restaurants know what you want to eat before you do. FOOD and DRINK INC. — MAGAZINE. (May 2016). https://www.bostonglobe.com/magazine/2016/05/26/how-restaurants-know-what-you-want-eat-before-you/hnZHM3xCkL1BhX0PKL3tmM/story.html
[12] Upserve. 2009. Upserve. (2009). https://upserve.com

A TRANSLATION

Restaurant related terms used and corresponding translation in terms of usage in this study.

- INGREDIENTS - any of the foods or substances that are combined to make a particular dish, this term is used to denote the data attributes in restaurant industry for big data
- CONSUME - eat, corresponds to gathering of big data
- RECIPE FOR SUCCESS - corresponds to dig data analytics
- KITCHEN TOOLS AND GADGETS - corresponds to solutions and tools available for big data application in restaurant industry
- FLAVORFUL IMPLEMENTATIONS - corresponds to real life big data implementation in the restaurant industry
- HELL'S KITCHEN - It's a popular reality television cooking competition show full of challenges, corresponds to challenges of using big data in restaurant industry
- Umami - Japanese food term to describe delicious food or taste
- POS - point of sales system to capture sales in the restaurant

Big Data Analysis in Finance Sector

Dhanya Mathew
Indiana University
711 N Park Ave
Bloomington, Indiana 47408
dhmathew@iu.edu

ABSTRACT

Big data as the name implies, refers to large and complex data which continues to grow enormously day by day. The broad proliferation of data and new and efficient technological support has transformed the way industries operate and compete. Industries like financial firms, in particular, have widely adopted big data analytics to obtain better investment decisions with consistent growth. In order to understand what drives profit in an organization or company, we should be able to predict the business trends, challenges, opportunities risks and what profit group (extremely unprofitable, average, extremely profitable etc.) a set of customers falls into based on their data at any given time. Financial firms like banks are storing these data for many decades and the recent technology boom that happened with big data technologies help the firms to uncover the secrets to understand consumer behavior, prevent major disasters and theft. We show the wide possibilities open for financial firms by analyzing big data to improve decision making, productivity, customer satisfaction etc which in turn beneficial for both organizations and customers.

KEYWORDS

i523, HID328, big data, data-driven, data lakes, Hadoop, Random Forest

1 INTRODUCTION

There are 3 fundamental elements to big data - Volume, Variety and Velocity. Data is stored and analyzed at a speed which is nothing but the velocity [8]. Data is getting increasingly gathered by low-cost and innumerable information-sensing Internet of Things devices like radio-frequency identification (RFID) readers, aerial (remote sensing), wireless sensor networks, cameras, software logs, microphones etc and hence causing data sets to grow rapidly [15]. Ben Walker of Voucher Cloud came up with a big data info graphic in 2015. According to Ben, the data generation per day is around 2.5 Quintillion Bytes and it would measure the height of 4 Eiffel Towers if it was stored in Blu-ray discs stacked on one another. Ben suggests that data generation by 2018 will be 50,000 GB per second [10].

According to Gartner Survey held in 2013, 64 percent of organizations were planned to invest or already invested in big data technology (including but not limited to Hadoop, NoSQL, Spark, R and Storm) [12]. Recent survey research indicates that 71 percent of firms in the financial services industry at a global level are exploring big data and predictive analytics [9]. This number continues to grow and sectors like government, business, technology, universities, health-care, finance, manufacturing etc make use of big data to obtain meaningful information using big data technologies [15]. We investigate in particular, how big data is helpful in financial firms in terms of predictive analysis and profitable growth. The finance sector also contributes to the daily data generation from products and marketing, banking, business, share market etc. Finance is a very sensitive field and any useful insight can make a positive impact on the overall turnover. Historic data analysis and real time data analysis are equally important in terms of finance sector. The key idea behind is how to retrieve the 'signal' of relevant information form the bulk of data. Let us explore the wide range of possibilities of big data analysis that finance sector can come up with including decision making, discovery of new business opportunities, enhanced productivity and efficiency, risk management, fraud detection, innovation possibilities, efficiency and growth and customer segmentation.

1.1 Efficient Decision Making

The era of big data helps financial firms to take quality business decisions related to expanding revenues, managing costs, hiring resources etc. based on effective data analysis which provide access to real-time insights. Data-driven decision making is one of the key advantages of big data technologies. Data driven decision making approach includes data storage, data elaboration, data analysis and decision making [9].

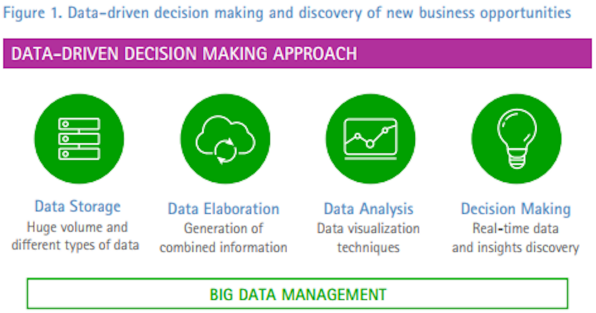

Figure 1: Data-driven decision making approach [9]

Figure 1 shows data-driven decision making approach and discovery of new business opportunities.

Data Storage: Even though big data does not define by the size alone, we need the right means to store the huge volume and variety of data. Big data is distributed - stored across many machines and managed with Hadoop File System and distributed DBs like HBase and Apache Cassandra [1].

Data Elaboration: Generate combined information by eliminating unwanted data using data cleansing methods like grouping, joining, filtering etc.(Spark, R, MapReduce, Storm).

Data Analysis: Big data analysis is the process of analyzing the data to derive the semantics of the available data to understand the hidden patterns, correlations, market trends, customer preferences which helps the organizations to take more informed decisions. Visualization tools include- Tableau, Google chart, D3, Fusion chart etc. are used to visualize the results of analysis.

Decision Making: Data-driven decision making based on the analysis.

There is a feedback analysis done for a bank as part of 2nd International Symposium on Cloud Computing and Big Data. For an organization, feedback processes are very important. This will help the organization to recognize the potential areas of improvement and also to identify gaps in services offered if done at regular intervals. This unnamed bank also participated in a feedback process. They collected the feedback over a period of 3 years and 6 months. They gathered the data from customers visited the bank branches and online users. Customers were given a feedback survey form. They were asked to rate on a scale of 1 to 5 on the below parameters. They could do this anonymously.

- Whether he/she is satisfied with the services quality?
- Whether customer is satisfied with the turn around time?
- Whether the customer queries are effectively addressed?

An analysis was performed on the feedback collected from around 20,000 customers. This is actually only a subset of the actual data collected [11].

As shown in Figure 2, the bank got an average rating on services offered. After that, Bank took some drastic measures to rectify the issues. This resulted in improvements in customer ratings.

1.2 Increased Productivity and Growth

Compared to traditional data warehouses, the big data concept of Data lakes to store raw data offers more flexibility in data access and analysis. Large volumes of data are stored, managed and analyzed in data lakes by using automated and sophisticated analytical tools.

Applications like, Machine learning algorithms, In-memory technologies, fast access DBs, big data queries and real-time analysis methods consume less time to come up with meaningful information and reports by accessing data lakes.

Data Lakes: Data Lakes can be compared to the actual lakes where rivers or streams that bring water to it. In data lakes, this is called ingestion of data. We collect all the data that we require to analyze to reach our goal irrespective of the source. These 'streams' of data come in several formats: structured data (simply said, data from a traditional relational database or even spreadsheet: rows and columns), unstructured data (social, video, email, text etc.), data from all sorts of logs (weblogs, clickstream analysis etc.), XML, machine-to-machine, IoT and sensor data. Logs and XML are also called semi-structured data. There can be data filters in place based on the requirements [3].

1.3 Fraud Detection

One of the best ways to fight cybercrime is with early detection. Banks are prime targets for cybercriminals and fraudsters, and any kind of public breach creates a lot of embarrassment, bad publicity, and unwanted scrutiny. Clearly banks have a vested interest in any technology to identify and prevent a data breach or fraud [5].

Financial institutions use analytics to identify fraudulent transactions from the genuine ones. Analysts can identify normal behavior based on the customer's past transactions. By applying analytics and machine learning, they can easily identify a fraud transaction based on the unusual behavior. An analysis system can have automated responses such a fraudulent transactions, including blocking that particular transaction. This stops the fraud even before it occurs. This can improve customer satisfaction and profitability of the bank [4].

A Security and Fraud analysis was done as part of the 2nd International Symposium on Big Data and Cloud Computing. Fraud analysis coupled with behavior analysis with past transactions and customers consumption capacity will reveal a potential threat to bank and as well as uncover past frauds [11].

As shown in Figure 3, the credit card transactions per card are increasing with time and the net ratio with previous month is the same. From Figure 4, we can notice that, in the month of May and June 2013, card number ending 13 shows a spike in transaction count. The transactions are doubled during the said period for this particular card. Ideally, an analyst should sound an alarm in this particular case. When we upscale to include millions of customers, such spikes are dangerous. This means a potential system compromise. This clearly indicates a misuse of the card and unauthorized access of funds by frauds.

1.4 Customer Segmentation and Personalized Marketing

Customer segmentation helps banks to transform from product-centric to customer-centric business. Big data enables the bank to group customers into segments. Customer segments are derived from the data sets. The dataset includes demographics, transactions, interactions with online and telephone customer services. And also external data, such as housing price. Financial institutions can then run targetted campaigns based on this segments [4].

There are many segmentation identification algorithms available in the Big Data world. Random Forest is one of the prominent algorithm. Apache spark, R are some of the technologies that have good integration with segmentation algorithms

Personalized Marketing: Using big data technologies, financial services firms can analyze their customer's merchant records and social media profiles to get a complete picture of their needs. This kind of marketing is done primarily by understanding customer's individual buying habits. It is beyond segment-based marketing and is called personalized marketing. Once those needs are understood, big data analysis can create a credit risk assessment in order to decide whether or not to go ahead with a transaction [4].

1.5 Understand New Business Opportunities

Big data will essentially change the manner in which business operate and compete. Institutions that are heavily invested in the big data space will have an identifiable advantage over others. As more and more data is generated, a performance gap will continue to grow. Emerging technologies (enable faster and easier data

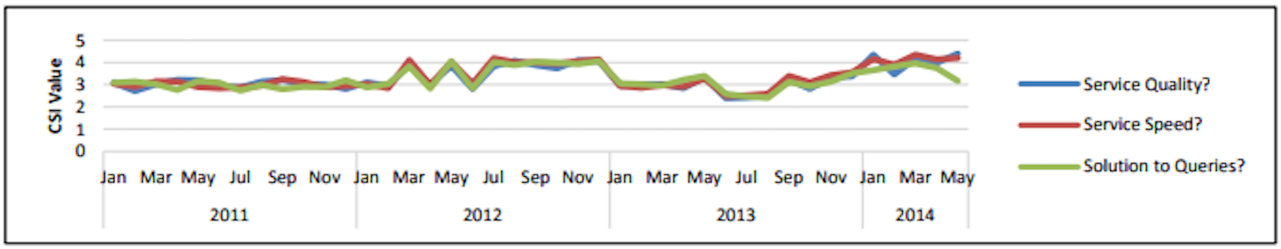

Figure 2: Overall Customer Feedback for provided parameters [11]

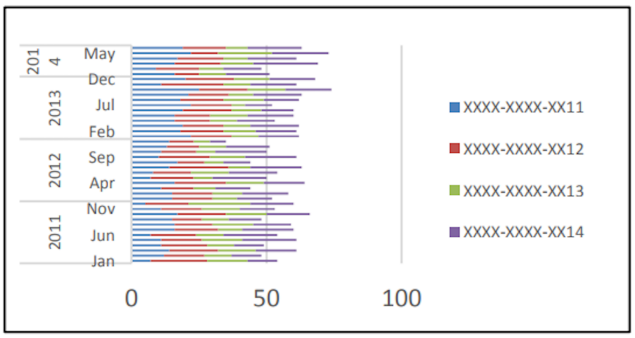

Figure 3: Net credit transactions count [11]

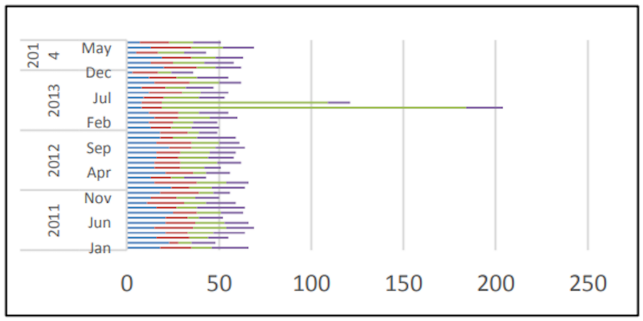

Figure 4: Net debit transactions count [11]

analysis) and digital channels will help them to offer better solutions. It is difficult to identify what is most important in the data, which technologies best suits the needs, who the customers are and what they expect. Being more data-driven gives an edge over competitors [2].

Big data incorporated with data science and business strategy can provide significant competitive advantages to organizations by offering new business opportunities. It allows companies to equip the data with pertinent real-time information when making decisions in order to eliminate inefficient operating processes, enhance the customer experience, take advantage of new markets, etc. For many companies and businesses, big data is already a critical path to develop new products, services and business models [9].

1.6 Discovery of Innovation Possibilities

Data analysis is increasingly becoming a key differentiator between wildly profitable and struggling businesses. Exploring and analyzing data, translates information into insight and drives to innovations [13].

Successful firms make decisions based on facts and data rather than intuition and are open to innovation concepts.

1.7 Risk Management

Financial firms especially banking sector are facing new regulatory requirements and challenges or risks each year. Big data adoption provides organizations a simplified and data-driven solution to mitigate the risks and helps to convert the data into usable information for regulatory reporting. Using data lakes and stronger analytic tools also helps to foresee the expected impact quickly [14].

1.8 Cost Effective Information Gathering

Unlike traditional business intelligence systems, new techniques and technologies used with Big Data allow to gain useful information at a much lower cost. New architectures and the move from data silos to 'data lakes' can provide substantial cost advantages and greater scalability due to flexibility in the data analysis. In fact, having all data sources in a data lake allows users to pull new reports on relatively new data, while in traditional data warehouses (DWHs) users have to extract, transform and load (ETL) new data into a static data model, which is expensive and costly from a time perspective. By using automated and sophisticated analytical tools that can store and analyze data faster and more easily, organizations can reduce the overall cost [9].

1.9 Big data - Risks and Considerations

Big data plays an progressively important role in the financial services sector. It is used for every single thing from targeting advertisements to optimizing portfolios. While these technologies have many benefits, critics are quick to point out that they can also become a source of discrimination if they are developed and/or used in an improper way [6].

Data Security: When we are considering the logistics of data collection and analysis, concern about data security is obvious and often takes top place in our mind. Data theft is an uncontrolled and growing area of crime. Security-related attacks are getting wider and more damaging [7].

Data Privacy: Data privacy is closely related to the issue of data security. We need to ensure that people's personal data are safe from criminals. Also, it is essential to be sure that the sensitive information being collected and stored are not going to be misused by yourself or by people who are authorized to analyze and report on it [7].

Bad Analytics: Bad analytics means 'getting it wrong'. It is obviously a risk to misinterpret the data patterns shown by the data where in fact there may simply a random coincidence. For example, product sales may increase following a major sporting event. Sales data will show this rise prompting you to relate your product with sports fans. But in fact, the rise was because of more people in town. The same situation can happen after a big music event as well [7].

Bad Data: There can be scenarios where data projects start off by collecting immaterial, erroneous, or out of date data which tend to have less time in designing the project strategy [7].

2 CONCLUSION

The Big Data revolution offers new opportunities for profitable growth and the financial services firms being one of the most risk-laden and dynamic of all business segments globally, are responding to it enthusiastically. It has become their derived knowledge that making sizeable investments in big data is ultimately a gain. Data has become the key element for decision making with the right choice of analytical tools and skill-set. When data from multiple sources combined and analyzed in a smart way, there emerges the insights which derive intelligent decisions and finally drives to profit.

ACKNOWLEDGMENTS

The author would like to thank the web loaded with information on any subject. The author would also like to thank Prof. Gregor von Laszewski for his review and suggestions.

REFERENCES

[1] Antony Adshead. 2013. Big data storage: Defining big data and the type of storage it needs. (April 2013). http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs

[2] EY. 2014. Big data Changing the way businesses compete and operate. (April 2014). http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf

[3] i scoop. 2016. Data lakes and big data analytics: the what, why and how of data lakes. Web page. (2016). https://www.i-scoop.eu/big-data-action-value-context/data-lakes/

[4] ingrammicroadvisor. 2017. 5 Big Data Use Cases in Banking and Financial Services. Web page. (Feb 2017). http://www.ingrammicroadvisor.com/data-center/5-big-data-use-cases-in-banking-and-financial-services

[5] ingrammicroadvisor. 2017. The Top 5 Trends for Big Data in Financial Services. Web page. (Sep 2017). http://www.ingrammicroadvisor.com/data-center/the-top-5-trends-for-big-data-in-financial-services

[6] Justin Kuepper. 2017. The Problem With Big Data in Financial Services. (January 2017). http://www.investopedia.com/articles/insights/010517/problem-big-data-financial-services.asp

[7] Bernard Marr. 2015. The 5 Biggest Risks of Big Data. (June 2015). http://data-informed.com/the-5-biggest-risks-of-big-data/

[8] Trevir Nath. 2015. How Big Data Has Changed Finance. (April 2015). http://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp

[9] Fabrizio Sarrocco, Vincenzo Morabito, and Gregor Meyer. 2016. Exploring Next Generation Financial Services: The Big Data Revolution. (2016). https://www.accenture.com/t20170314T051509__w__/nl-en/_acnmedia/PDF-20/Accenture-Next-Generation-Financial.pdf

[10] Storage Servers. 2016. How much data is created daily? Web page. (Feb 2016). https://storageservers.wordpress.com/2016/02/06/how-much-data-is-created-daily/

[11] Utkarsh Srivastava and Santosh Gopalkrishnan. 2015. Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. *Procedia Computer Science* 50, 1 (May 2015), Pages 643–652. https://doi.org/10.1016/j.procs.2015.04.098

[12] Conn STAMFORD. 2013. Gartner Survey Reveals That 64 Percent of Organizations Have Invested or Plan to Invest in Big Data in 2013. Web page. (Sep 2013). http://www.gartner.com/newsroom/id/2593815

[13] tableau. 2017. Big data. Web page. (2017). https://www.tableau.com/solutions/big-data?utm_campaign=Prospecting-BGDATA-ALL-ALL&utm_medium=Paid+Search&utm_source=Bing&utm_language=EN&utm_country=USCA&kw=%2Bbig%20%2Bdata&adgroup=CTX-Big+Data-Sitelink&adused=%7bcreative%7d&matchtype=p&placement=%7bplacement%7d&gclid=CNGAtcXn09YCFWwifgodDW4Dsw&gclsrc=ds&dclid=CMPCu8Xn09YCFcjVZAodCT4PlQ

[14] David Turner, Michael Schroeck, and Rebecca Shockley. 2013. Analytics: The real-world use of big data in financial services. (May 2013). https://www-935.ibm.com/services/multimedia/Analytics_The_real_world_use_of_big_data_in_Financial_services_Mai_2013.pdf

[15] Wiki. 2017. Big data. Web page. (Oct 2017). https://en.wikipedia.org/wiki/Big_data

Big Data Analytics and Edge Computing

Arnav Arnav
Indiana University, Bloomington
Bloomington, Indiana, USA
aarnav@iu.edu

ABSTRACT

With the exponential increase in the number of connected IoT devices, the data generated by these devices has grown enormously. Sending this data to a centralized server or cloud results in enormous network traffic and may lead to failures and increased latency. A solution for this problem is to do some processing on the devices closer to the network edge, enabling responsive and real-time analytics. There have been various developments in the field of edge computing some of which are described here.

KEYWORDS

i523, HID201, Edge Computing, Big Data Analytics, IoT, Platforms

1 INTRODUCTION

Internet of things is rapidly gaining importance and Evans Data Corporationfis Global Developer Population and Demographics Study reports 6.2 million developers working in the IoT domain [11]. With the rapid increase in the acceptance of Internet of Things (IoT) devices across various fields across the world, ranging from industrial sensors to lifestyle and sports products, and the consequent increase in the data generated by such devices, there is a pressing demand for devices and processes that can analyze this data and provide responsive analytics [13]. Traditionally, IoT applications follow one of the two approaches - "cloud-centric approach, where the sensing devices send data to the cloud where the analytics are performed or device-centric approach, where stand-alone devices running proprietary code perform analytics locally" [13]. Networks are largely centralized with organizations soring all data, which may not be directly beneficial to them, in their data centers, and data flowing from the edge to the cloud on each operation [1].

With an increase in the number of connected devices, it gets increasingly difficult to perform all analytics on a server in a traditional manner. Thus, edge computing involves pushing a part of this computation closer to the end user of the device, or closer to the network edge [14][1]. This helps reduce the cost incurred in communicating large amounts of data over the network, ensures some level of availability even when the connection to the cloud is broken and reduces the cost of computation and storing data on the cloud [13][1].

2 HOW EDGE COMPUTING WORKS

Edge computing emerged with the development of content delivery networks (CDNs) by Akamai which use nodes close to the user to prefetch web content and accelerate web throughput. Edge computing extends this concept with the help of cloud infrastructure to run arbitrary task-specific code at nodes close to the edge, typically known as cloudlets. These cloudlets usually run on a virtual machine or a light-weight container for ease of isolation and resource management [12].

Proximity to the edge of the network ensures various benefits. It helps to provide highly responsive applications, by using a more powerful computing resource near the edge and minimizing end-to-end latency, which is essential in time-critical applications like virtual reality which require a latency of less than 16ms for the images to appear stable [2][12]. Proximity also increases scalability with the help of edge analytics where cloudlets perform the first level of analytics on the sensor data and only send processed data and metadata to the cloud to reduce bandwidth usage as the number of connected devices increases [12]. Decentralization of data can also provide the owners of data more control over the privacy of their data, and provide ways to safely communicate this data between various entities [1][4].

In industrial applications like aviation where a large amount of data is generated on each flight [12], analyzing this data in a centralized manner becomes impractical. In such cases, fog computing is more useful which adds different elements at various levels of hierarchy between the edge and the cloud [6]. In industrial environments, there are a lot of different systems running new as well as legacy applications which may be proprietary and integrating these applications to provide end-to-end IoT solutions is still a challenge. Linux Foundation's EdgeX platform provides a way to simplify and standardize edge computing architectures and is gaining importance as an industrial IoT solution [6].

3 SOME EXAMPLES

Simmhan describes an application that was built using Apache-NiFi, a lightweight dataflow execution engine used for vehicle classification from video streams using a "Tensorflow deep neural network encapsulated within a NiFi dataflow executing across multiple raspberry pis" [12]. This allows video streams to be analyzed locally and also provides the flexibility to use cloud infrastructure for computation when edge devices are constrained [13].

Yang Zhao et al proposed an occupancy and activity monitoring application with doppler sensing and edge analytics. The application uses low-cost motion sensing and embedded signal processing, detection and machine learning to detect activity in real time, even when multiple people are present in a room. The developers provide a web portal to help ease monitoring activity from a remote location [16].

Analysing video feeds on a large scale in real time is a challenging task. Each of the videos may be very large and a large amount of bandwidth is needed to stream the video feed to a central location which is not feasible especially if the cameras are connected wirelessly. In addition to this, the entire video may not be useful and most parts of it may be discarded depending on the application.

Furthermore, these applications need to provide results with low latency as important decisions often need to be made based on the output in case of surveillance applications [2]. Thus compute abilities available on cameras can be utilized to provide real-time video analysis, processing the video at the camera and only communicating interesting bits to the cloud [12].

A real-time video processing solution is proposed in [2] that focuses on traffic planning and safety and provide high accuracy outputs and detects anomalous traffic patterns to suggest preemptive safety measures and reduce traffic accidents and deaths. Interactive augmented reality applications must rely on object tracking, face detection, and other video analytics to obtain spatial knowledge, and must rely on cloudlet based edge solution to provide users with a smooth interaction experience [2].

Scientists at MIT's Computer Science and Artificial Intelligence Lab (CSAIL) are working on self-folding printed robots and their use in saving lives as an alternative to invasive surgery procedures, which would require a cloud in the proximity as those robots and sensors generate a large amount of data that needs to be processed very fast [3].

Verizon created a universal cloud-in-a-box solution running Linux on a generic x86 architecture, in an OpenStack container that can put compute, storage and networking resources near the edge to support their increasing number of users and power 5G in the future [3][8].

4 RECENT DEVELOPMENTS

The need for a holistic data analytics platform to combine various techniques in cloud and edge computing and to ease data management arises in applications like health monitoring where anomalies in a patient's conditions must be immediately reported and an analysis on historical patient data is needed to find out more details about patient's overall condition [10]. The lack of platforms for the edge that allow development and simple deployment in a distributed setting is a key limitation to using edge devices effectively [13].

Taking on these challenges a serverless platform was proposed by Nastic et al that supports real-time analytics across cloud and the edge by optimizing the placement of analytics operations and automatically managing available resources [10]. The model takes a top-down approach for control processes combined with a bottom-up approach for data management allowing analytics to be done at various levels of granularity and the results served to the application from either the edge or the cloud depending on the need [10]. The platform provides developers with an API that allows them to easily define analytics functions without worrying about data management and optimization complexities.

Early IoT infrastructures were heavily cloud dependent and all the computation was done in the cloud. This tight coupling with the cloud is however not desirable in many time-critical or data-intensive applications [4]. An edge offloading architecture named FADES (Function virtualizAtion basED System)was proposed by Cozzolino et al that reverses the traditional paradigm and dispatches some computation to the devices close to the edge. How this offloading to the edge should be performed depends on the application, the hardware capabilities, and the software requirements [4]. The multilayer pipeline ensures reduced amount of data to be uploaded and the MirageOS based unikernel approach provides an additional layer of security by running the deployed tasks inside a virtualization platform, bridging the gap between complex cloud-based applications and edge applications and providing modularity [4].

5 AI ON THE EDGE

With the emergence of decentralized applications, smart machines that rely on machine learning and mesh computing to provide local real-time analytics are becoming a reality. MIT's Eyeriss which is an accelerator for deep neural networks uses no wifi and no data transmission. With peer to peer networks gaining importance, edge computing is vital to provide low latency applications that are decentralized [1].

Since many artificial intelligence (AI) applications need a huge amount of processing power and require a large amount of data, traditional AI applications rely on cloud servers to perform their computation. This is a serious limitation in applications where connectivity is not reliable and time-critical decisions are required [5]. iEx.ec is a company that uses Etherium blockchain to create a market for computing resources, in turn, facilitating distributed machine learning [7].

In applications like flying a swarm of drones, a loss of connectivity to the cloud can be fatal and cause disruption of the operation. Thus AI coprocessor chips that can run machine learning algorithms can offer intelligence at the edge devices. Movidius recently announced a deep learning compute stick [9] that can add machine learning capabilities to computers and raspberry pis as a plug and play device [5].

Machine learning algorithms line one-shot learning which require lesser data are rapidly enabling edge devices to perform intelligent tasks easily [15]. "Gamalon, backed by Defense Advanced Research Projects Agency (DARPA), is using Bayesian Program Synthesis to reduce the amount of data required for machine learning" [5].

6 CONCLUSION

With the increase in the number of connected devices and the increase in the demand of real-time and interactive applications, we see that edge computing is a necessity and many industries are rapidly moving towards edge solutions. Although industrial IoT still faces challenges with the integration of legacy applications and proprietary applications with new technology, open source solutions are being widely accepted. Research on various platforms and architectures for edge computing continuously aims to reduce the gap between cloud and edge devices and establish standards for the same. The emergence of decentralized applications and the growing importance of machine learning has driven technologies that provide machine learning capabilities to edge devices which are becoming a fundamental requirement to move towards decentralized AI applications, that can provide results in near real time.

ACKNOWLEDGMENTS

The author would like to thank Professor Gregor von Laszewski for providing the opportunity to study the topic in detail and for

providing all the tutorials and support material needed to write the paper.

The author would also like to thank other associate instructors of the class for helping promptly with queries on piazza which helped everyone a great deal.

REFERENCES

[1] Scott Amyx. 2016. Ready for the disruption from edge computing? IBM iot blog. (August 2016). https://www.ibm.com/blogs/internet-of-things/edge-computing/

[2] G. Ananthanarayanan, P. Bahl, P. Bodk, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha. 2017. Real-Time Video Analytics: The Killer App for Edge Computing. *Computer* 50, 10 (2017), 58–67. https://doi.org/10.1109/MC.2017.3641638

[3] Jason Baker. 2017. Why OpenStack is living on the edge. opensource.com blog. (May 2017). https://opensource.com/article/17/5/openstack-summit-news

[4] Vittorio Cozzolino, Aaron Yi Ding, and Jörg Ott. 2017. FADES: Fine-Grained Edge Offloading with Unikernels. In *Proceedings of the Workshop on Hot Topics in Container Networking and Networked Systems (HotConNet '17)*. ACM, New York, NY, USA, 36–41. https://doi.org/10.1145/3094405.3094412

[5] Ben Dickson. 2017. How do you bring artificial intelligence from the cloud to the edge? TNW website. (August 2017). https://thenextweb.com/contributors/2017/08/21/bring-artificial-intelligence-cloud-edge/#.tnw_5VcrJGrz

[6] Andrew Foster. 2017. Why the Industrial IoT Needs an Open-Source Edge Platform. (july 2017). https://www.rtinsights.com/why-the-industrial-iot-needs-an-open-source-edge-platform/

[7] iEx.ec. 2017. *Building a Fully Distributed Cloud for Blockchain based Distributed Applications.* white paper. iEx.ec. http://iex.ec/wp-content/uploads/2017/04/iExec-WPv2.0-English.pdf

[8] Nicole Martinelli. 2017. Pushing the edges with OpenStack. Open Stack Articles. (May 2017). http://superuser.openstack.org/articles/edge-computing-verizon-openstack/

[9] movidious.com. 2017. movidius deep learning stick. movidius website. (2017). https://developer.movidius.com/

[10] S. Nastic, T. Rausch, O. Scekic, S. Dustdar, M. Gusev, B. Koteska, M. Kostoska, B. Jakimovski, S. Ristov, and R. Prodan. 2017. A Serverless Real-Time Data Analytics Platform for Edge Computing. *IEEE Internet Computing* 21, 4 (2017), 64–71. https://doi.org/10.1109/MIC.2017.2911430

[11] Avi Patwardhan. 2016. Incorporate streaming analytics in the Internet of Things. IBM data and analytics hub blog. (october 2016). http://www.ibmbigdatahub.com/blog/incorporate-streaming-analytics-internet-things

[12] Mahadev Satyanarayanan. 2017. The emergence of edge computing. *Computer* 50, 1 (2017), 30–39. http://elijah.cs.cmu.edu/DOCS/satya-edge2016.pdf

[13] Yogesh Simmhan. 2017. IoT Analytics Across Edge and Cloud Platforms. IEEE IOT Newsletter. (May 2017). https://iot.ieee.org/newsletter/may-2017/iot-analytics-across-edge-and-cloud-platforms.html

[14] Wikipedia. 2017. Edge computing — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=Edge_computing&oldid=802381553 [Online; accessed 7-October-2017].

[15] Wikipedia. 2017. One-shot learning — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/w/index.php?title=One-shot_learning&oldid=793877024 [Online; accessed 7-October-2017].

[16] Yang Zhao, Jeff Ashe, David Toledano, Brandon Good, Li Zhang, and Adam McCann. 2016. Occupancy and Activity Monitoring with Doppler Sensing and Edge Analytics: Demo Abstract. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM (SenSys '16)*. ACM, New York, NY, USA, 322–323. https://doi.org/10.1145/2994551.2996543

Big Data and Massive Online Open Education

Weipeng Yang
School of Education, Indiana University Bloomington
201 N Rose Ave
Bloomington, Indiana 47405
yang306@umail.iu.edu

ABSTRACT

Massive Open Online Education(MOOC) often refers to a kind of online course that emphasizes free and unrestricted access via the internet. The research will focus on how to collect such data as length of study, learning time of the day, preferred courses and other learners' behaviors to analyze and make predictions for future course designers and online teaching platform administrators.

KEYWORDS

i523, HID236, big data, MOOC

1 INTRODUCTION

Nowadays people in the field from K-12 to institutions of higher education have witnessed a great shift from traditional classroom teaching to distance learning. These online courses are based on Learning Management Systems (LMSs). The more advanced a learning management system could be, the better they could collect a variety of data from students' activities and performances.= [5]. With this amount of data, educators and researchers could utilize data mining and data visualization techniques to generated synchronous feedback and to keep-track of students' progress with higher efficiency. The source of the data could be the average time a student is learning a certain course, test score of a quiz, response rate in an online discussion session, etc [8].

2 BIG DATA AND EDUCATION

Big data are sometimes considered as the amount of data that could be available to an organization, in which contains a gigantic amount of information that would render itself too complex for any household data-processing software to manage. Not only big data emphasizes the amount of data, it also all calls for the need of real-time data and a wide source of data. As for the field of education, big data provide more direct evidence-based approach to learning and could allow researchers to see the difference of students nationwide or even worldwide. Analyzing these data could help LMS providers to make strategical improvements or let online teachers provide personalized tutoring to those student in need [1].

3 EDUCATIONAL DATA MINING IN MOOC

Educational Data Mining (EDM) refers to the process of analyzing various kinds of data on diverse levels of education (for instance, data could be collected from online students to decision-makers of the MOOC corporation) via a variety of techniques and tools [3]. What is collected includes time, sequence and context of the course being taught. It could be easily inferred that EDM requires cooperation of various subjects such as statistics, artificial intelligence, machine learning, etc. With these interdisciplinary means data could be processed. EDM aims to predict the academic performance of a student, evaluate student learning in the context of LMS, improve instructional sequences and evaluate add-on software that could provide additional help with the LMS. Pioneers of this field are researching how to improve the modeling of student performance, teaching domain and LMS properties and characteristics. They are also interested in providing students with diverse needs different track of learning courses accordingly [2]. It is concluded that EDM usually requires five means to analyze educational data:

(1) Prediction
(2) Clustering
(3) Relationship Mining
(4) Distillation of Data
(5) Discovery via Models

Prediction stresses that students' academic performance will be analyzed via their behavior in online learning. Clustering means that by sensing such specific characteristics as preference of learning materials or performance styles, the students will be grouped according to the elements mentioned above. Moreover, these resources could be recommended to learners with similar needs. Relationship mining is the most mentioned method in EDM. It focuses on figuring out hidden relationships with such variables as teaching and learning strategies, students' performance in online environment and students' interactions. Distillation of data for human judgement concentrate on ways to filter out most important data in a cluster so that researcher could figure out structures in the data quickly. Discovery via models, the last method, focuses on utilizing existing model to analyzed newly collected data [2]. To better utilize EDM, the MOOC cooperation shall establish a data structure first by determining the need of its users and their learning goals as well as the source of the data. Then they shall start defining certain variables and start creating a model or choose from an existing one. In the end they could start using this model to predict students' preferences and make modification accordingly. It is stated in some research that after using EDM to collect more information, MOOC's learning outcomes are improved and course tutors could cater to students' needs more efficiently [3].

4 LEARNING ANALYTICS

Other than computer science and statistics, Learning Analytics (LAs) are rooted on a wider spectrum of subjects such as sociology and psychology. Those who applies LAs wish to create a learning environment for teachers in which each student's learning need will be satisfied to the greatest extent and they could also choose their own learning tracks via their own learning habits. LAs could also enable facilitator of MOOC to distribute educational resources with better decision-making mechanism, providing feedback to students

and help at-risk students (refer to those who haven't participated in learning activities for a long time or those who do not have ideal performance in quizzes) [6]. To reach these goals mentioned above, LAs also involve a vast variety of data, from students' learning habits, assignments collected, social interaction online, threads on discussion forums to generate students' progress and identify those who might be at-risks. With enough data collected, LAs could also be utilized to determine the overall structure of the course, students' learning objectives and the sequence of learning contents [3]. Decisions are often based on models with multiple dimensions such as students' experience, knowledge, their preferred sequence of learning. LAs consist of three steps: Data collection and processing, analyzing data and action and data post processing. It can be concluded that LAs could be efficient for all level of users, tutors and decision-makers in MOOC however it also face lots of challenges as decision makers will need to determine what kind of data to collect and they need to connect separate collected data together via specific algorithm to gain a holistic view of inner connections between.= [12].

5 DISCUSSIONS

Although the application of big data in MOOC will bring convenience and efficiency to online teaching and learning, challenges are also lurking within the boundaries of ethical, researching and equality issues.

5.1 Ethical issues

Not only in the field of online education, the ethical aspect is a heated debate topic almost across any place that involves application of big data. For instance, privacy factor is perhaps the most discussed item about the era of big data [9]. As mentioned above, learning management system developers are designing course recommendation systems like online shopping websites. Website may push such messages to students as "You scored well in this subject. According to our analysis, those who excel in this course are more likely to take this" or "You have taken these three courses in our row. We would recommend this course to you since it is in the same cohort." This does seem promising, however in real world such scenario might take place: Based on the pretest and history performance of a student, the system suggest that he/she is most likely to drop out of this course. Instead of supporting him/her to master this course (which might cost extra manpower), the staff of the education website might intentionally let him/her fail, only charging their tuition fee[11]. One other issue is that with the implication of such recommendation system, students are more likely to follow the track set by the algorithm and the element of serendipity is deprived. Therefore, they cannot explore the field of knowledge freely [10].

5.2 Issues in Research

Although research could collect a greater variety of data nowadays in MOOC, it could not depict the full nature of online teaching and learning. Ergo the data we could collect determines the research directions and endeavors of MOOC. Researchers would build algorithm to study the pattern of online teaching, yet the data researchers failed to collect will become the missing variables or coefficients in the algorithm thus brings biased outcomes. Moreover, researching MOOC data requires a team with interdisciplinary backgrounds and cooperation of multiple organizations and institutions. It would take lots of negotiations and agreements to set up such cooperation [4].

5.3 Equality Issues

MOOC data are collected from those who have access to internet which means that those who reside in area with underdeveloped cyberinfrastructure are not likely to be counted in the research [7]. On the other hand, some data are collected by organizations for profit and required certain amount of fee to retrieve. Thus, some researchers or non-profit organization may not have access to them. Such scenario is against the openness spirit in MOOC.

6 CONCLUSIONS

With the rapid iteration of software and LMSs, learners nowadays could easily gain access to massive amount of learning materials at almost every corner of the world. MOOC brings great flexibility to learners so that they could choose their online instructors, their sequence of learning and their learning materials. It also brought instructors and students closer than ever so that they could interact more frequently, thus enabling a more dynamic atmosphere in learning. Challenges also arises as the numbers of students grow exponentially, it could be more difficult for teachers and managers to keep track of each student's status and provide help accordingly. Thanks to the introduction of big data in this field, decision maker of MOOC could evaluate and investigate students' status more easily and could develop more learning strategies. It can be envisioned that soon big data will relieve more burden on teachers' shoulders.

REFERENCES

[1] Terry Anderson. 2008. *Theory and practice of online learning.* Edmonton, AB, Canada: AU Press., Chapter Teaching in an online learning context, 343–366.

[2] Ryan Baker and Kalina Yacef. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining* (2009).

[3] Matthew Berland, Ryan Baker, and Paulo Blikstein. 2014. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning* (2014).

[4] D. Boyd and K. Crawford. 2012. Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon, Information. *Communication and Society* (2012).

[5] Dave Cormier and George Siemens. 2010. The open course: Through the open door: Open courses as research, learning and engagement. *Educause Review* (2010).

[6] H. Drachler and G. Wolfgang. 2012. Confidence in Learning Analytics. In *Learning Analytics and Knowledge 2012 Conference, Vancouver, Canada, April 29-May 2.*

[7] R. Eynon and A. Geniets. 2012. *On the Periphery? Understanding Low and Discontinued Internet Use Amongst Young People in Britain.* Technical Report. Nominet Trust, Oxford, UK.

[8] Jane Klobas and Tanya McGill. 2010. The role of involvement in learning management system success. *Journal of Computing in Higher Education* (2010).

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* McKinsey: McKinsey Global Institute.

[10] K. Robins and F. Webster. 1989. *The Technical Fix: Education, Computers and Industry.* New York: Macmillan.

[11] M. Spector. 2013. Emerging educational technologies: Tensions and synergy. *Journal of King Saud University-Computer and Information Sciences* (2013).

[12] M. Spector, M. Merrill, and J. Elen. 2014. *Handbook of research on educational communications and technology.* New York, NY: Springer.

Big Data Analytics in Higher Education Marketing

Ashley Miller
Indiana University
admille@iu.edu

ABSTRACT

While the collection of vast amounts of data in the world of higher education has occurred for decades, the use of big data applications and analytics is fairly new to this environment. There is a need to understand how the use of big data analytics can help institutions determine student behavior as well as stay relevant in a digital and evolving age of technological advances, tools, and skills. The higher education space is changing as the population of students going to college is on the decline which increases competition and the need for institutions to be more strategic in their efforts for attracting students to their institutions. We will explore at a very high level how higher education could utilize big data analytics to inform marketing initiatives in recruiting and enrolling students as well as what potential challenges and considerations could impact this process.

KEYWORDS

i523, hid329, big data, higher education, marketing, analytics, data-driven decision making

1 INTRODUCTION

Today's colleges and universities are drowning in data [2]. With the emergence of big data, institutions are now faced with providing useful analysis and reports to a variety of stakeholders including administrators, professors, as well as to the students themselves [2]. A variety of challenges lie in the path of institutions using big data effectively such as finding the necessary skill set for staff, technology tools and resources, as well as understanding then what to do with the data collected to better inform decision making. While there is literature that addresses utilizing big data for learning analytics and even course enrollment and development, as Daniel states, there is still "limited research into big data in higher education" [2]. Higher education could benefit from using big data analytics in their marketing efforts for recruiting and enrolling students as well as identifying what gaps may still exist in the quest to understand today's college student in their college search process.

2 CURRENT ENVIRONMENT

According to the *Western Interstate Commission for Higher Education (WICHE)*, the projected number of high school graduates will decline over the course of the next decade [1]. Meanwhile, the number of four-year institutions in the United States has increased with more than 3,000 available college options [3]. Increased competition and fewer students have made the higher education marketplace crowded and convoluted. There are a variety of factors that go into a student's decision on where to attend and ultimately what area to study. In their 2013 trends report, the Lawlor group identified a number of aspects that will impact the higher education landscape, among those included are [8]:

- The demographics of today's college student is changing with more women attending college than men in addition to an increase in ethnic and socio-economic diversity as well as first-generation students [1].
- The college search process today happens primarily in the digital space which includes third-party websites, email, social media, and digital advertising [4]. This *Generation Z* grew up in a technology rich and connected environment which means that colleges have to also be constantly on in this space to effectively recruit and enroll students [4].
- The need to showcase the *value* of going to college, not only through the quality of education received relative to the price paid but also through outcomes-level data, including retention and placement rates and even starting salaries of recent graduates.

With these trends in mind, there is a need for institutions to be more targeted in their marketing efforts. Big data analytics can be used to help assess the impact of these trends as well as how institutions can make better decisions with these techniques.

3 BIG DATA ANALYTICS TO SEGMENT BY DEMOGRAPHICS

Big data can be one way to better inform these efforts and also help with the return-on-investment (ROI) for advertising and marketing related efforts. Other universities have capitalized on utilizing big data in attracting students. For instance, St. Louis University described a process of retroactively looking at demographics of students who succeeded at the university and had high satisfaction scores [9]. This information coupled with nearly 100 other data points gave insight to the admissions team when exploring new markets as well as identified clusters of students that may be interested in attending St. Louis University [9]. The university was then able to develop a targeted digital campaign in these areas that they believed included students who would be a good fit for the university. With the reliance on big data, St. Louis University was able to reduce costs as the need to mass market went away and ultimately increased enrollment and retention rates [9].

4 BIG DATA ANALYTICS TO UNDERSTAND BEHAVIOR ONLINE

The web environment is common tool in college exploration as a report by Ruffalo Noel Levitz shows that three out of four high school students utilize an institution's website as their most used resource when exploring colleges [4]. Web analytics provides a wealth of information on users such as how much time is spent on certain pages, bounce rate, paths in website exploration and ultimately conversion rates when various goals are completed such as scheduling a tour or filling out an application for college [6]. Google Analytics is one tool used to track and evaluate efforts

on websites. Higher education institutions could take advantage of this tool by tracking top pages viewed, geography and age of visitors, as well as areas where they may be losing students in the information search process. With this data, institutions can identify opportunities for improvement in ensuring students are finding the information they need in a timely and efficient way as well as develop customized marketing efforts to invite students back into the experience to complete various calls-to-action.

5 BIG DATA ANALYTICS TO CONVEY VALUE

Utilizing big data to understand the outcomes of current students, and ultimately graduates, can help tell the value story to prospective students [8]. By tracking the experiences among current students during their four (or more) year college career, predictive analytics could be implemented to determine which combination set of experiences best contribute to the success of a student. Temple University utilized predictive analytics to increase graduation rates by sending messages to students who were considered to be "at risk for dropping out" based on financial aid data [12]. This similar type of approach could be utilized in marketing efforts as well. If a profile of a student could be created based on existing data and therefore create an ability to predict the future actions of prospective students, then marketing messages could be more tailored based on where that prospective student is in the enrollment funnel.

6 CHALLENGES

In order for the use of big data analytics to be successful in higher education marketing, there are basic measures that have to be met. Marsh et. al outlines some key considerations when using big data analytics for effective decision making which include: accessibility, quality, timeliness, and motivation to use [5]. These same factors can be also impact the use of big data analytics in a higher education setting.

6.1 Accessibility

Typically, institutional research offices have been the primary house for student data collected over time, but that doesn't mean it's the only place where data lives [7]. As Daniel state, there is also data in higher education that lives across a number of areas in a wide variety of formats [2]. With this, accessing data can be a challenge as there is no central system or warehouse. Depending on the type of data needed, sources can live in silos which means that the data sources are not connected to one another to provide a complete picture. Further, the level of permission to access data can also vary which can make it difficult for marketers to access.

6.2 Quality

Coupled with the fact that data across an institution can live in multiple places, there are issues around the quality of data [2]. The disparity of data sources can lead to quality concerns but also the skill set of those who maintain or utilize the data. If no standard processes exist for data cleaning, integration, reporting, or interpretation, then the risk of having invalid conclusions increases [5]. Decisions made on inaccurate data could potentially be costly for institutions.

6.3 Timeliness

There can be issues with timeliness in a variety of ways. Alignment on the objectives for data analysis can require input from multiple stakeholders which takes time. The aspects involved in processing the data itself could involve a significant amount of time, people, and resources. Often times, decision making for marketing purposes needs to happen quickly and there can be a gap between obtaining the needed information and when decisions need to be made [5].

6.4 Motivation

There is also an underlying cultural aspect to using big data analytics in the right way across an institution. With the silos that exist in higher education, collaborating across departments and sharing information overall can help to forge better working relationships. Successful efforts rely on the involvement of multiple departments including information technology (IT) [2]. The importance and message about utilizing big data analytics has to come from leadership for others to be equally motivated.

7 POSSIBLE SOLUTIONS

While significant challenges can exist in utilizing big data analytics to inform marketing initiatives in higher education, there are possible solutions to explore. One way to overcome the challenge of accessibility would be to create a central area where data could live. This would also allow the opportunity for others to access data and create consistency across the institution. Having a central system would also help with the data quality aspect if the format of the data was consistent in the way it was stored, presented, and accessed. Along with creating a central area, a standardized data flow would also be beneficial. In Figure 1, Eduventures outlines a proposed data flow within the area of higher education [11].

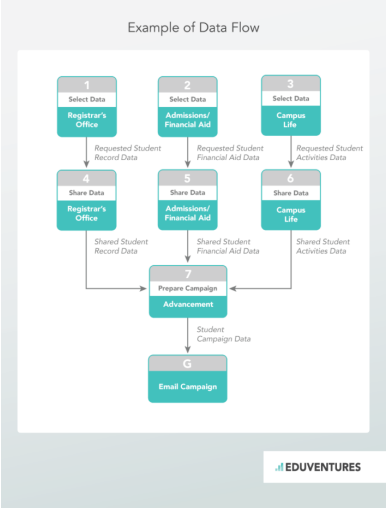

Figure 1: Example of a data flow [11]

8 OTHER CONSIDERATIONS

Throughout this process of exploring the use of big data analytics for higher education marketing, there are other factors to consider.

With the collection, analysis, and use of big data, what implications does this pose to data security and privacy issues among students? As stated by Slade and Prinsloo, ethical issues can come into play regarding data ownership and governance [10]. Given that higher education institutions are faced with an increased level of scrutiny, what protocols have to be put into place to ensure the safety of students' data? Further, what level of accountability is assigned with the different areas/persons that are in need of the data to inform decision making? There are also policy issues to consider regarding what kind of data can be collected on students and how and where this information should be stored.

9 CONCLUSIONS

Competition for today's student will only increase with changing educational needs and offerings, including the development of emerging degree programs as well as delivery. For marketers in higher education, they need to have access to necessary data about current as well as prospective students to better tailor messaging and marketing efforts appropriately. With this, the validity of available data is key as making decisions based on incomplete data can be problematic and costly for an institution. Given the nature of the web environment that is constantly changing, obtaining data in a timely manner is crucial so action can be taken at the right time. Insights around data are only as good as the people that make use of them, so creating a culture within an institution that motivates others to make data-driven decisions is imperative for these efforts to be successful.

10 ACKNOWLEDGEMENTS

The author would like to thank Gregor von Laszewski at Indiana University Bloomington for his instruction and feedback as well as Juliette Zerick for providing a first review and recommendations for improvement.

REFERENCES

[1] Peace Bransberger. 2017. Impact and Implications: Projections of Male & Female High School Graduates. (2017).
[2] Ben Daniel. 2015. Big Data and Analytics in Higher Education: Opportunities and Challenges. *British Educational Research Association* 46, 5 (2015), 904–920.
[3] National Center for Education Statistics. 2015. Digest of Educations Statistics. (2015). https://nces.ed.gov/fastfacts/display.asp?id=84
[4] Stephanie Geyer. 2016. E-Expectations Trend Report. (2016). https://www.ruffalonl.com/papers-research-higher-education-fundraising/recruitment-marketing-and-financial-aid/e-expectations-research-reports
[5] Julie A. Marsh, John F. Pane, and Lara S. Hamilton. 2006. Making Sense of Data-Driven Decision Making in Education. (2006). www.rand.org
[6] Mohammad Amin Omidvar, Vahid Reza Mirabi, and Narjes Shokry. 2011. Analyzing the Impact of Visitors on Page Views with Google Analytics. *International Journal of Web and Semantic Technology* 2, 1 (2011), 14–32.
[7] Anthony G. Picciano. 2012. The Evolution of Big Data and Learning in Analytics in American Higher Education. *Journal of Asynchronous Learning Networks, Volume 13: Issue 3* 16 (2012), 9–20. Issue 3.
[8] Hanover Research. 2014. Trends in Higher Education Marketing, Recruiting, and Technology. http://www.hanoverresearch.com/media/Trends-in-Higher-Education-Marketing-Recruitment-and-Technology-2.pdf. (2014).
[9] Jeffrey Selingo. 2017. How Colleges Use Big Data to Target the Students They Want. (2017). https://www.theatlantic.com/education/archive/2017/04/how-colleges-find-their-students/522516/
[10] Sharon Slade and Paul Prinsloo. 2013. Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist* 57 (2013), 1510–1529. Issue 10.
[11] James Wiley. 2016. Do Your Know Where Your Data is Going? (2016). http://www.eduventures.com/2016/09/where-is-your-data-going/
[12] Mikhail Zinshteyn. 2016. Big Data Allows for Higher Education Predictive Analytics. (2016).

3

Big Data Applications in Electric Power Distribution

Swargam, Prashanth
Indiana University Bloomington
107 S Indiana Ave
Bloomington, Indiana 47408
pswargam@iu.edu

ABSTRACT

Now-a-days, the process of storing the power measurements have changed. Conventional meters are replaced by the smart meters. New distribution management systems like SCADA and AMI are implemented to monitor power distribution. These smart meters record the readings and communicate the data to the server. However, these systems are designed to generate the readings very frequently i.e., 15 minutes to an hour. Upon that, smart meters are being deployed at every possible location to improve the accuracy of the data. This advancements in electric power distribution system results in enormous amounts of data which requires advance analytics to process, analyse and store data. Implementation of Big Data technologies, challenges of implementing Big Data in Electric Power Distribution Systems,Architectures used in implementation are discussed here.

KEYWORDS

HID228, I523, Big Data, Power Distribution,Smart Power

1 INTRODUCTION

Volume of data is increasing. According to Ref [12], it is said that, world's data utilization will increase to 44 zettabytes from the current utilization of 4.4 zettabytes. To process this data, Big Data analytics will be useful. But, instantiating a big data architecture is not easy task.

In electrical Power Distribution industry, data deluge is picking its pace. The data which was recorded for month, is now being noted for very small intervals. This quadruples the amount of data that should be process. There is a lot of potential work to be put in for designing a good Big Data architecture to process and analyse this data. Most of the power generation units are developing their infrastructure to support these designs.

1.1 4 v's in Big Data in Power Distribution System

Big Data is mostly described in 4 v's. Each of this V's are considerable factors in a Big Data Solution [3].

Volume: The data is periodically generated by many data sources like smart meters, machines and other appliances.

Variety: Each data source in electric power distribution system is explicit to each other. Each source has its own frequency of data generation and its own method of data generation. Thus, the data is heterogenous.

Velocity: is the speed at which the data is available for the end user.

Veracity: It deals with the correctness of the data. As all the data collected by sensors, meter tend to have various losses, correction algorithms should be defined to find the accurate data. Their might be chances for data transfer losses.

2 DATA SOURCES

Smart meters which are placed at customer's vicinity will record the consumption of a specific group of customers. This data can be used to analyse the behaviour of customer for certain circumstances of weather and environment.

Distribution systems which manage the distribution of power, generate large amount of data related to voltages and currents at various levels of distribution. This data is very important in analysing the load level and demand for the distribution circle [2].

Phasor measuring units at generation. This data is used to analyse the behaviour of generator and amount of power generation that will be required to supply enough power. This data will be used to decide the functioning of generators [10].

Old market data will be used to analyse the pricing and marketing strategies. These data is more focused on users and their behaviour.

3 DATA INTEGRATION

3.1 Service Oriented Model

This model has a workflow which is defined in Business process Enterprise Language often referred as WS BPEL [6]. WS BPEL is used for is enterprise language used for automating a business process. BPEL files defines the process to be followed by a request from the web services. In this model, All the user requests are handled by services. These services either connect to the storage resources or calls the other services based on the process model defined in BPEL. This modelling ensures data is being utilised in a structural manner and analysed according to the process model.

Interfacing services: This service is used to manage the interfaces with the end user. This services generally initiates calls to a process defined in WS BPEL. After all the other processes which are defined in process model are completed, this service is used to project the analytical data to the user at the end of execution. In this case, this service receives data from one of the process models [7].

Execution Service: This service is responsible for all the logic involved in modelling the data. For the common requests, these are well documented in BPEL files. These documents specify the set of instructions to be followed to model the data as per the request from the service. This service uses a Information management services to establish a data link to data storages [7].

Pooling Services: All the data requests coming from Information management services are managed by pooling services. This service help the other services in establishing a dynamic connection to data storages. This service also handles one way communication between the data storages and Information management services.

This is called event-driven approach. All the activities like addition of data, removal of data in data storages are considered as event. This events are communicated to the information management services.

4 DATA STORAGE AND PROCESSING

4.1 Hadoop and MapReduce

Hadoop and MapReduce are prevalent technologies in storing and processing data. Hadoop has a database in file system called as HDFS [4]. HDFS and MapReduce is an Apache Project which is used to split the data into various segments and store the data in various commodity boxes. This boxes are clustered together to allow the flow of data between them.

As the data is generated at different physical locations, it will be easy to store data at different geographical locations. There will be minimal transmission of data. Changes in electrical grid doesn't require the change in entire data model. On addition or deletion of a electrical node, a new data storage can be added without any intervention to the existing data storages. This distributed model also ensures high availability. Availability of one data source will have minimal or no effect on the availability of the system thus reducing the downtime and business losses.

The data from various sources have different formats. This makes it difficult to store data in traditional relational databases because of type conversions and relational handling. Hadoop overcomes this problem by storing the data in filesystems. Data can be easily pre-processed and stored in the pictorial representations rather than in tables and schemas.

Mapreduce is a programming model. This has two components i.e., Jobtracker and Tasktracker. Jobtracker is a master process which is responsible for scheduling assigning the jobs to Tasktracker. Tasktracker is responsible for execution of the mapreduce jobs.A sample mapreduce task takes has two phases [11]. The first phase is a map phase, where the data is divided into several pieces. The second stage is reduce phase, where the data is processed to produce output.These mapreduce jobs are scheduled and run in batches.This is called Batch Processing.

This map and reduce functions are very reliable in analysing the nature and demand of customer from the data available from the most recent processed jobs. Mapreduce jobs run on static data. This will not serve the requests like load analysis, electrical machinery failure, metering failure, power loss which require real time data [8].

4.2 Apache Spark for Realtime data

Apache Spark is a cluster computing model. It has capability to perform real time analysis of data. IT is nurtured with more enhanced machine learning algorithm and libraries [14]. Spark SQL, MLlib, Spark streaming, GraphX are some of those. Spark framework contains data in distributed sets. It also has set of working programs on the distributed sets of data. This set of programs are called Resilient Distributed Dataset functions [9].

The dynamics of electrical properties changes in milliseconds. In order, to collect these dynamics, the power measuring systems have evolved . New instruments like phasor measurement units have evolved. These devices collect data at the rate of 20-40 readings per second. However, if there is any delay in processing such huge amount of data, then the collected data is not useful. Apache spark tackles this issue in two different approaches.

Streaming Approach: Streaming approach reacts to the each and every event that occurs in the data. As soon as new data is injected, all the resilient distributed dataset functions are called. This function processes data and makes them into a usable format and stores them [14]. This kind of approach is used in metering, billing and load management.

Iterative approach: In this approach, spark offers in memory computing. The datasets are accessed in memory instead of the going to the physical database [5]. All the phasor readings which are required by multiple requests to calculate state space estimation use the developed cache data on the servers instead of accessing them from the data storage. This make requests like state space calculation much lighter.

5 CHALLENGES IN IMPLEMENTING BIG DATA

5.1 Information Security

A large amount of customer electricity usage data is collected. This data must be protected from data leaks. Access control systems must be enhanced to restrict the access to the customer data. Leaked data can be exploited to trace the end user and his/her appliances [13].

5.2 Asset Management

Assets are the power collection units. These are one of the important devices in the architecture. All the assets must be maintained properly to ensure the quality of data. If any of the power measuring unit goes down or malfunctions, there will be discrepancy in analysing data. This will lead to improper decisions.

5.3 Adaptability

The amount of data in increasing by many folds. In present world, Data Analytics has become a part of Electrical Industry. Though, Many Power Industries have implemented Big Data solutions, there are many industries which are yet to implement Big Data technologies. Most of the South asian countries still use SCADA for processing electrical Data [1].

6 CONCLUSION

This brief description highlights the advancements of Big Data Solutions in Power distribution systems. Firstly, Data sources for analytic systems in power distribution like smart meters, Phasor measurement units are briefed. Integration of Data from various sources using service oriented architecture and the important processes in the service oriented architecture are discussed. Later, Implementation of distributed file system i.e., HDFS with processing models like MapReduce and Apache Spark are discussed. At last, challenges like information security, asset management and adaptability of Big Data Technologies are discussed .

REFERENCES

[1] ABB.COM. 2002. ABB SCADA system to automate power for Hyderabad & Secunderabad and streamline APCPDCL electrical distribution network. (2002).

http://www.abb.com/cawp/seitp202/bdaf43d0073a9eb965256cb60021c734.aspx
[2] A. B. M. Shawkat Ali (Ed.). 2013. *Smart Grids Oppurtunities,Developments, and Trends*. Springer, School of Information and Communication Technology, Central Queensland University, North Rockhampton, QLD Australia.
[3] Amr A.Munshi and Yasser A.-R.I. Mohamed. 2017. Electric Power Research Systems. *Elsevier* 151 (2017), 68–85.
[4] Apache Hadoop. 2017. Apache Hadoop. (2017). https://en.wikipedia.org/wiki/Apache-Hadoop
[5] Justin Kestelyn. 2013. Putting Spark to Use: Fast In-Memory Computing for Your Big Data Applications. (11 2013). https://blog.cloudera.com/blog/2013/11/putting-spark-to-use-fast-in-memory-computing-for-your-big-data-applications/
[6] MANAGEMENT MANIA. 2015. WS-BPEL (Web Services Business Process Execution Language). (2015). https://managementmania.com/en/ws-bpel-web-services-business-process-execution-language
[7] Jyotishman Pathak, Yuan Li, Vasant Honavar, and James McCalley. 2007. A Service-oriented Architecture for Electric Power Transmission System Asset Management. In *Proceedings of the 4th International Conference on Service-oriented Computing (ICSOC'06)*. Springer-Verlag, Berlin, Heidelberg, 26–37. http://dl.acm.org/citation.cfm?id=2170265.2170269
[8] Shyam R, Bharathi Ganesh HB, 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 2015), 171–178. SMART GRID TECHNOLOGIES.
[9] S Sagiroglu, R Terzi, Y Canbay, and I Colak. 2016. Big data issues in smart grid systems. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*. 1007–1012. https://doi.org/10.1109/ICRERA.2016.7884486
[10] Abu-Rub Shady S. Refaat, Haitham, Rub, and Mohamed Amira. 2016. Big Data Better Energy Management and Control Decisions for Distribution Systems in Smart Grid. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, DC, USA, 1–6. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7840966
[11] Tutorialspoint. 2017. Hadoop - MapReduce. (2017 2017). https://www.tutorialspoint.com/hadoop/hadoop-mapreduce.htm
[12] EMC Digital Universe with Research & Analysis by IDC. 2014. The Digital Universe of Oppurtunities: The Rich Data and Increasing value of Internet of Things. (04 2014). https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
[13] N Yu, S Shah, R Johnson, R Sherick, M Hong, and K Loparo. 2015. Big data analytics in power distribution systems. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. 1–5. https://doi.org/10.1109/ISGT.2015.7131868
[14] Matei Zaharia. 2017. Apache Spark. (2017). https://en.wikipedia.org/wiki/Apache-Spark

Big Data Analytics in Weather Forecasting

Himani Bhatt
Indiana University
Bloomington, Indiana
himbhatt@iu.edu

ABSTRACT

Big data analytics applications help data scientists, statisticians to analyze tons of data in order to have useful insights and make predictions based on it. In majority of the cases the data that has been dealt with is unstructured and complex. In case of weather forecasting, data about the current state of atmosphere (that includes wind, humidity, temperature etc) is gathered, and by using concepts of meteorology, the evolution of atmosphere in future is predicted.

KEYWORDS

HID 202, i523, NWP(Numerical Weather Predictions), Supercomputers.

1 INTRODUCTION

Daily over 20 terabytes of data get generated of different key weather-related parameters like wind speeds, temperature readings, satellite images, barometric pressure from different locations. Yet the extreme weather events cause huge amount of loss and damage worldwide. For instance, 90 percent of the crop losses are due to weather calamities all over the world. The ability to better predict the weather could slash this by up to 25 percent. The ability to use the big data generated efficiently directly affects the chances to reduce economic loss, environmental damage and fatality. This can be considered as the best motivation to improve and invest in this extremely challenging task.Weather forecasting is considered extremely challenging since large number of variables are involved and their interaction is also very complex. The data accumulation is done from the sources like trained observers, weather balloons, weather stations, satellites, radar, etc. The data is then plugged into super computers which uses numerical forecast equations to create forecast models of the atmosphere and to produce the meteorological analysis. [7].

2 HISTORY OF WEATHER PREDICTION

After the invention of the first electronic computer ENIAC, a group of meteorologists at New Jersey's institute for Advanced Study produced first weather forecast using ENIAC and numeric prediction techniques back in 1950. There forecast was for 24 hours but they took more than 24 hours to complete. This can be marked as the need for the beginning of numerical weather predictions [10].

Numerical weather prediction models

Any typical numeric weather prediction model will have complex and multidimensional data. These models divide earth into various atmospheric boxes. For each box they try to apply mathematical equations for the current weather and they try to forecast weather

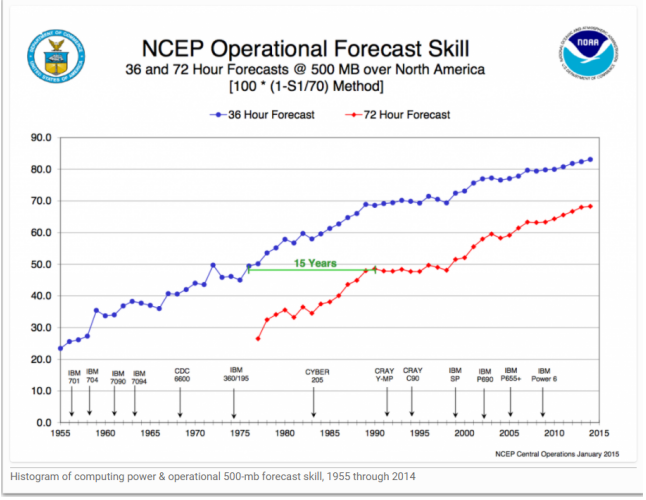

Figure 1: Histogram by NCEP(National Centres for Environmental Prediction) showing trends in computing power and operational model skill over time [10].

for coming days. These equations are derived from physics and fluid motion. This data is divided horizontally as latitude longitude and vertically as different pressure levels. And hence it is multidimensional.

The higher the resolution of the data, more is the accuracy of the NWP models. Because of improved resolution, improved data sources, and improved physics processes, enormous advancements are seen in the atmospheric modelling techniques. With the models improving, requirement of high computing powers increase proportionally, and the models now need supercomputers to run operationally.

3 WEATHER PREDICTION IN MODERN TIMES

A company named Weather Analytics has been forecasting a week's weather in advance by analyzing US Government's weather service data of past 38 years. It's methodology includes placing global data in a 35km by 35km grid and then extraction of relevant variables from the generated outputs which are often in range of 650,000 per hour.

The company creates, calculate and extract relevant variables. This huge data with weather information is stored in databases, from which data presentation, data packaging, and calculation of

additional variables is done [2].

Enhancement of data processing power is clearly exemplified by the supercomputer at UK based firm called Met Office, which has 480,000 cores, two million GB of memory and 17,000TB of storage. This 140 tonnes processing giant is capable of performing over 23,000 trillion operations per second, which has brought the accuracy of modern 4 days forecasts on par with that of 1 day forecasts 30 years ago through improved data model resolution and enabling usage of smaller grid sizes [2].

4 HADOOP FOR CLIMATE ANALYTICS

The NASA center for climate simulation uses Apache Hadoop for performing high performance computing as it combines distributed storage of large data sets with parallel computing and optimizes computer clusters. It has built a new platform with Hadoop for developing new analysis capabilities. Hadoop Bloom filter is utilized that helps to identify rapidly and memory efficiently if an element is present.

Advantage of using HDFS and MapReduce

Hadoop is resilient to failure, provide load balancing and parallelization. When the data is sent to an individual node, it is also sent to other nodes in the cluster. So in case of failure their will be a copy of data available.Storage nodes and compute nodes are same. It means that tools for the data processing are on the same servers where the data is located, which result in fast data processing. Hadoop is capable of processing terabytes of unstructured data in just minutes and petabytes in hours. Thus is highly used in weather analytics.The requested operations are mapped to the appropriate nodes using specified key [8].

5 CURRENT WEATHER FORECASTING MODELS

5.1 Deep Thunder'- World's Most Advanced Hyper-Local Weather Forecasting Model

Deep thunder is a research project by IBM and it is headed by Lloyd Treinish. The scientists in IBM developed first parallel processing supercomputers that can be used for weather modeling. This supercomputer is based on IBM RS/6000 SP(it is a family of Reduced instruction set computer based on UNIX servers, supercomputers made by IBM in 1990's). It was first installed at National Weather Service office in Georgia in 1966.
With high accuracy deep thunder can deliver hyper localized weather conditions up to three days in advance, with calculations as fine as every mile and as granular as every 10 minutes. Deep thunder can forecast weather for an 84 hours duration.Rio De Janeiro's city operation center is already using Deep Thunder [1].

Deep thunder uses a 3D telescopic grid where data from one model is fed into another model, that is called coupled models in climatology. This data is then verified with the historical data. They work in collaboration with National Oceanic and Atmospheric Administration(NOAA), and use global models provided by them.

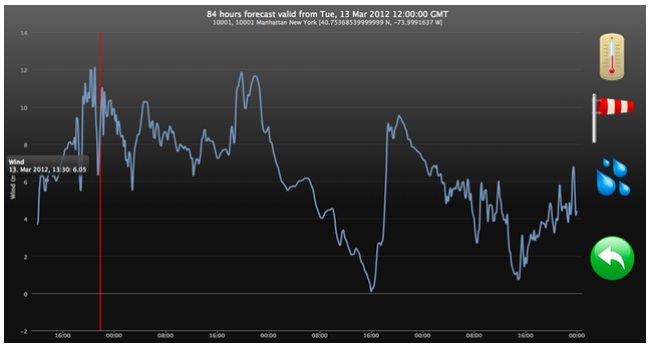

Figure 2: 84 hours forecast by predicted by Deep Thunder [3].

In order to decrease the amount of processing required, they zoom in, that exponentially increase the resolution, down to models with resolution as small as 1 meter. This layered model approach shrinks big high performance computing problem to a relatively smaller footprint of parallel processing systems. The data sources of Deep Thunder are public satellite sources, underground personal weather stations, smart phone barometer etc. [5].

5.2 Hybrid renewable energy forecasting(HyRef)

Together with focusing on the damages done by weather we can also work on creating renewable resources. Wind and solar energy can be used to produce enormous amount of power. Forecasting the wind and solar availability will help in reducing the uncertainty associated with variable renewable energy generation. Hyref combines big data analytics and weather modeling technology to accurately forecast the availability of wind power and solar energy. This helps the energy providers to optimize the output and integrate more renewable energy into the power grid [4].

Hyref has the following key components :

- **Weather modeling capabilities** - The modeling is done at very granular level, from a square kilometer to vertical dimensions like heights where rotors and turbine hubs are located.
- **Advanced imaging technology** - Cloud imaging, advanced cameras.
- **Sensors on the turbines** - Highly perceptive sensors are used to monitor the turbulence, temperature and direction of wind.
- **Analytical capabilities** - Hyref use cloud image analytics and advanced numerical prediction models to calculate weather impacts on solar generation and to forecast cloud movements. SAS is used for this purpose and it is on DB2 platform [6].

According to Brad Gammons, general manager of IBM's Global Energy and Utilities Industry group, development of an intelligent

system obtained by the combination of weather and power forecasting will increase the system availability and will optimize power grid performance. He believes that power of analytics and big data will help in tackling the intermittent nature of renewable energy and forecast power production from solar and wind, with the efficiency never experienced before [9].

6 CONCLUSION

Ability to better predict the weather can have a dramatic impact on our planet by creating new energy resources and helps us to be better prepared for weather related incidents across all industries. Businesses from retail, production, energy, agriculture, water resource management are all using predictions from weather models to make decisions. Thus a better weather forecast will definitely have a positive economic impact. Not only economic impact, it can potentially save thousands of lives and safeguard property in times to come.

Increasing evidence of climate change worldwide is prompting governments and scientists to take action to protect people and property from its effects. One such instance is the up-gradation of national weather information system by South Korea, after being hit by Typhoon Sanba and Hwangsa storms. The upgrade has increased agency's data capacity drastically. And it has now become Korea's most capable storage system. The output comes as the better understanding of the weather patterns and predicting the ferocity and location of the weather events.This project of South Korea dramatically illustrates today's big data phenomenon and its impact on weather forecasting.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the Assistant Instructors for their feedback and help.

REFERENCES

[1] David Semmelroth Alan Anderson. 2015. BIG DATA AND WEATHER FORECASTING. (2015). http://www.dummies.com/programming/big-data/data-science/big-data-and-weather-forecasting/

[2] Manek Dubash. 2016. Big data and the weather forecast. (2016). http://www.zdnet.com/article/big-data-and-the-weather-forecast/

[3] SEAN GALLAGHER. 2012. How IBMfis Deep Thunder delivers fihyper-localfi forecasts 3-1/2 days out. (2012). https://arstechnica.com/information-technology/2012/03/how-ibms-deep-thunder-delivers-hyper-local-forecasts-3-12-days-out/

[4] Rolf Gibbels. 2013. IBM's Hybrid Renewable Energy Forecaster. (2013). http://www.altenergymag.com/content.php?post_type=2129

[5] IBM. 2012. Deep Thunder. (2012). http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepthunder/

[6] IBM. 2013. IBM's HyRef Seeks to Solve Wind's Intermittency Problem. (2013). http://www.renewableenergyworld.com/articles/2013/08/ibms-hyref-seeks-to-solve-winds-intermittency-problem.html

[7] Social Media. 2013. Big Data: The Force Behind Weather Modeling. (2013). https://www.youtube.com/watch?v=nGi4xtyyY_c

[8] NASA. 2015. Applying Apache Hadoop to NASA's Big Climate Data. (2015). http://events.linuxfoundation.org/sites/events/files/slides/ApacheCon_NASA_Hadoop.pdf

[9] Michael Graham Richard. 2013. Better short-time local weather prediction cheaper renewable energy. (2013). https://www.treehugger.com/renewable-energy/better-short-time-weather-prediction-cheaper-renewable-energy.html

[10] Chris Robbins. 2015. A Brief History of Weather Forecasting. (2015). http://www.iweathernet.com/educational/history-weather-forecasting

Big Data Analytics in Agriculture

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

ABSTRACT

The modern agricultural industry faces numerous challenges. The global population is increasing rapidly. As the global population grows the agricultural industry must find a ways to increase to the production out of each and every acre in order to match the growing need. Many parts of the world face food insecurity issues. Food often spoils during food transport as a result of inefficient food packaging or problems with food transportation. This results in substantial food waste and potential health hazards. Big Data and data science is now being adopted into the agricultural industry to help solve some of these issues. The necessary technology is becoming increasing available and is increasing affordable. The use of wireless technology, the Internet of Things, smart devices, and sensors is becoming commonplace throughout the world. Big data is being collected, analyzed, and delivered to back stakeholders to enable better management of operational activities. As a result, farming production processes are becoming more productive and food delivery systems are becoming more reliable.

KEYWORDS

I523, HID332, precision farming, smart farming, food production, food safety, precision agriculture, big data

1 INTRODUCTION

Big Data is revolutionizing the Agricultural Industry. The Internet of things together with the availability of cloud technology is creating a new phenomenon called Smart farming [7]. Big data is a term for datasets that are so large or complex that traditional data processing applications are not inadequate to process them [7]. Large amounts of information is being captured, analyzed, and used to make operational decisions [3]. As a result, farmers are optimizing productivity, reducing costs, reserving resources, and increasing profitability.

Big Data Analytics is also reducing waste and spoilage as food moves through the food supply chain. According to McKinsey and Company, approximately one-third of all food in lost or wasted every year. That equates to a nine hundred forty (940) billion dollar Global impact [5]. Much of this occurs during the food shipment process.

Internet connected devices are becoming common place on farms. Almost all new farm equipment has sensors. Sixty percent of farmers report some type of internet sourced data to make operational decisions [1]. Sensors are becoming common in food packaging. The related software market is growing rapidly. In 2010 the investment in Agricultural Technology was 500 million. In 2015 the investment had grown to 4.2 billion [3].

2 BIG DATA

Big data represents information assets characterized by such high volume, velocity and variety as to require a specific set of technology and analytical methods for its transformation [7]. The amount of data and information generated by the food production industry is massive. For example, it is estimated that sensors on harvesting equipment generate about seven gigabytes per acre. There are 93 million acres of corn and 80 acres of soybeans in the United State alone [1]. In India, there are one billion acres. Data is being collected at the micro-bit level and much of this data is being processed in real time [4].

3 THE SMART FARM AND PRECISION AGRICULTURE

3.1 Precision Agriculture - Overview

Precision agriculture is a specific farm management technique that uses sensor and analytic technology to measure, observe and respond to crop and livestock management in real time. Precision farming matches farming techniques to the specific crop and livestock needs. The objective of precision farming is to ensure that crops receive that exact inputs that they need, at the correct time, and in precise amounts [6]. Examples of crop inputs include: water, fertilizer, herbicides and pesticides. This strategy enables a farmer to get the most productivity out each and every resource. Solutions are customized to each individual farmers unique needs.

Processes that are typically managed with Precision techniques include: seeding, planting, harvesting, weed control [2], fertilizer management, breeding, disease control, pesticide management, light and energy management [7].

3.2 Precision Farming - Benefits

Precision farming techniques give farmers the ability to make operating decisions in real time based upon data and information that is being generated in real time. It also gives farmers the ability to make predictive insights in farming operations [7]. All of this results in significant benefits: Increased yields, reduced costs, greater productivity, immediate disease management [4], improved crop quality, and better cash flow. Big Data makes farms more profitable. Also, when inputs such as herbicides and pesticides are better managed, it helps the environment. Precision farming also has a socioeconomic impact worldwide because efficiency improvements can help to alleviate global food insecurity [7].

3.3 Precision Farming - Data Collection

A very common approach to collecting data is sensor technology. Sensor technologies measure and monitor data. Sensors register

and report deviations in real time. Sensors include devices that are located locally on the farm and external satellites.

Types of local sensors include: connected farming equipment (tractors, harvesters), chips planted into livestock [3], and drones. Examples of the types of data that may be collected via local sensors include: Rainfall and water measurements, crop health, livestock health, weather information, yield monitoring, and lighting and energy management [7]. Drones can collect aerial images of fields. Aerial field images can help to monitor crop health. [6]. Data is oftentimes collected in very precise detail. For example, information can be gathered for for each square meter of land or for every individual plant [2].

Data collected with local sensors is often supplemented with information other external sources such as satellites and the cloud. Data that may be collected via satellite and available in real time on the cloud includes: Weather and climate data (historical and real time), soil type analysis, market information, and livestock movements. Data collected from orbiting satellites can also be very granular and personalized [3]. For example, soil characteristics such as texture, organic matter, and fertility is collected to the meter at locations throughout the world [3].

3.4 Precision Farming - Data Analysis

After the data is collected it must me consolidated and analyzed. A significant amount of this support is being provided by machine supplier companies that have been servicing the farming industry for generations such as John Deere, DuPont Pioneer, and Monsanto [4]. Now, in addition to selling seeds and machinery, these companies are selling decision support and data science services [1].

Most of this support is in the form of software decision support technology. Companies collect information from individual farms, combine this information with data other sources, including their own databases, and apply statistical models and algorithms. Results and recommendations are delivered to each grower as personalized solutions. Examples of some potential solutions are: how far apart to place seeds based upon the field position, or what to do to better manage nitrogen levels in the soil [4].

These companies have developed and maintain massive databases of their own. DuPont Pioneer has mapped and has collected data on 20 million acres in the United States. Another company, Cropin, which provides support for farmers worldwide, including growers in extremely remote areas, has mapped over one billion acres globally. Cropin can provide data by individual farm, farm clusters, districts, states, and even countries (India) [4].

In addition to big companies, there are also public institutions that are involved with Big Data Applications. These include universities, the USDA, and the American Farm Bureau Federation. Their interest typically involves issues such as food safety, food security, and data privacy regulation [7].

3.5 Precision Farming - Infrastructure

After the data is analyzed it is downloaded from the cloud and made available to the farmers, typically through wireless technology devices. It may be downloaded to an farmers Ipad or computer in a tractor. Other information can be sent to Smart phones. By interacting with the Internet of Things farmers can manage operational activities from anywhere in the world [7]. Other devices are self automated. One such self automated technology is Variable rate technology (VRT). Variable rate technology is built into equipment such as irrigation systems, feeders, and milking devices [6]. These devices automatically operate in such a way as to deliver optimal results with no human intervention.

None of these processes can happen without the appropriate infrastructure to store, transmit, and transform the data. Typical Storage vehicles for this data are typically cloud based platforms, Hadhoop Distributed file system, cloud based data warehouses and hybrid systems. Data transfer is accomplished via wireless technology using cloud based platforms. Machine learning algorithms are typically used to transform and cleanse the data [7].

3.6 Precision Farming - Decision Making

Below are some examples of ways in which information provided by Big Data Analytics is providing farmers with the information that they need to make more informed decisions concerning their operations.

Following are some examples of technology in the world of crop science: Satellite systems and sensors can monitor the development of crops in detail. Individual plants can be monitored for nutrients, growth rate and health [5]. In this way, disease outbreaks can be recognized and addressed immediately [4]. Entire fields can be mapped with GPS coordinates to collect data concerning soil conditions and elevation. The data in analyzed using Algorithms and the data is sent back to an Ipad on the farmers tractor. The tablet then communicates with the tractor's planting mechanism telling it exactly where to place every seed [4]. This same technology can even tell if a single seed has been missed [1]. GPS units on tractors, combines, and trucks help determine the optimal usage of equipment [5].

Big Data technology also improves the field of Animal and livestock management. Milk cows are tagged with chips that monitor the health of the animal. Milking machines shut down when the animal is sick [3]. Sensors indicate when livestock are ready to inseminate or give birth [2]. Smart dairy farms are using robots to complete tasks such as feeding cows, cleaning barns, and milking cows [7].

Consolidated data can offer insights and information that has never before been possible. Big data companies can test and gather information about the effectiveness of different kinds of seeds across many different conditions, soil types, and climates. The origin of crop diseases can be identified quickly and efficiently with web searches similar to the way that flu epidemics are currently identified [1]. This will enable players to take corrective action quickly. Historical analytics can determine the best crops to plant [7].

4 FOOD SAFETY AND THE FOOD SUPPLY

Big Data not only impacts primary food production, it helps to improve the entire food supply chain [7]. According to the Food and Drug Administration,food waste equates to approximately 680 billion in industrial countries and 310 billion in developing countries annually [3]. A significant amount of this food waste occurs during

food transport. Big Data can help to address this issue in various ways. First, it can help to manage the logistics of transportation. For example, Big Data can help to insure that food is transported in the best weather conditions in developing countries. This helps to avoid issues such as trucks not be able to navigate muddy roads. Big data can also assist coordination needs between supplier, retailer, and consumer. For example, consumer demand can be tracked with customer loyalty cards or retailers data on shopping patterns. Coordinating food delivery with consumer need helps to minimize food waste [3].

Food spoilage can also be monitored during food transport. Inadequate packaging of food often results in food waste and food spoilage that can even result in life threatening food borne illnesses [5]. Packaging sensors can detect gases that is being emitted from food when it starts to spoil. RFID based traceability systems can monitor food as it moves through the supply system. Packaging integrity and freshness can be monitored in real time. Therefore, waste is reduced and food quality issues can be addressed as they occur [5].

5 CHALLENGES AND ISSUES

5.1 Developing Countries

The challenges in developing countries are unique. In order for Big Data to be successful there must be infrastructure. Technologies such as satellite imagery and weather monitoring may not be fully developed. Small farmers can not always afford specialized machinery. Farmers do not always have access to devices such as computers, tablets, or Ipads [3].

Such issues are starting to be addressed in some countries. For example, in Africa organizations are being formed which pool several farmers resources together. This enables better access to resources as well as educational information. Also, there are establish companies that are starting to invest and develop technologies around the world, such as CropIn and Monsanto.[3]. Mobile devices such as Smart phones are becoming more common and are starting to be used more widely to manage information. For example, in Tanzania 30000 farmers use mobile phones for business purposes such as contracts, loans and payments.[6].

5.2 United States

In the United States, machine suppliers in the form of big companies have played a big role in this evolution by developing decision support tools that provide information to better manage farms [4]. When individual farmers share their personal data with big companies such as John Deere and Monsanto it raises some significant unanswered questions and concerns. Is my personal data safe? Is my data secure? Who owns the data? Who will profit from the data? [1]. Even if it is assumed that the original data belongs to the individual farmers, there is still the question of who owns the data after it is consolidated. Furthermore, there is concern that the aggregated data could be used to for malicious intent such as manipulation of commodity markets [7].

For these reasons, there need to be clear and defined standards regarding issues of privacy, security, data ownership, and market speculation. Such standards are only in the beginning stages of development. Organizations who are currently working on the farmers behalf to develop these standards include: The American Farm Bureau Association, The Big Data Coalition and AgGateway. In the interim, farmers need to do their best to fully understand any contracts that they sign in which they agree to share data. [7].

6 CONCLUSION

Big data analytics are improving our food delivery system in ways that are beyond substantial. Information is being made available to stakeholders that has previously been impossible to obtain. Big data is being referred to as the most significant revolution in farming productivity since mechanization. Today, billions of people worldwide cope with undernourishment and alarming food shortages. Big Data is expected to make an impact food insecurity throughout the world as more farmers adopt these techniques. This technology will enable even small holder farmers to make full use of their productive potential. Big Data technology is making the food delivery system healthier, safer, and more efficient. Big Data in agriculture is here to stay.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Dan Bobloff. 2015. Big Data Comes to the Farm. Web page as Article. (Sept. 2015). https://www.businessinsider.com/big-data-and-farming-2015-8/
[2] CEMA. 2015. Farming goes digital: The 3rd Green Revolution. Web page as Newsletter. (May 2015). http://www.cema-agri.org/newsletterarticle/farming-goes-digital-3rd-green-revolution
[3] Nir Kshetri. 2016. *Big Data's Big Potential in Developing Economies.* CABI.
[4] Katherine Noyes. 2014. Cropping Up on Every Farm Big Data Technology. Web page as Article. (May 2014). http://fortune.com/2014/05/30/cropping-up-on-every-farm-big-data-technology/
[5] Tim Sparapani. 2017. How Big Data and Tech Will Improve Agriculture from Farm to Table. Web page as Article. (March 2017). https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-and-tech-will-improve-agriculture-from-farm-to-table/#4512e9f85989
[6] Wikipedia. 2017. Precision Farming. Web page. (Sept. 2017). https://en.wikipedia.org/wiki/Precisionagriculture
[7] Sjaak Wolfert. 2017. Big Data in Smart Farming - A review. *Agricultural Systems* 153 (May 2017), 69–80. http; /sciencedirect.com/science/article/pii/s0308521x16303754

Big Data in Oceanography

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

ABSTRACT

This paper will give an overview of how big data is used to collect and process data about our oceans in better ways to understand it and how to solve complex issues surrounding it.

KEYWORDS

hid346, i513

1 INTRODUCTION

Ever since man as first seen the ocean, it has always had an air of mystery to it. Even now in the age of information, we still barely know anything about it. This paper is to give an overview of how and with what we are collecting the data and the problems we face in that collection of data. In the hope that we can start to uncover the mysteries that hide below the surface. To collect this data there is a use of many different sensors. [1]

Acoustic Doppler Current Profiler: This tool measures the speed and direction of ocean currents using the principle of Doppler shift. Measuring currents is a fundamental practice of physical oceanographers.

Technologies for Ocean Acoustic Monitoring: This technology listens to the ocean for all sounds, boats, sea animals, waves, siezmic activity.

The Bushmaster and the Chimneymaster: A collection net used to grab tube worms or living fauna near geothermal vents. These tools are typically attached to a submersible vessel.

Clod Cards: These plaster cards track the motion of water for benthic organisms. The organisms that inhabit the bottom of the ocean. allowing us to learn more about the harder to reach parts of the ocean.

Drifters: Drifters, are essentially devices that flow with the current of the ocean. Allowing for them to be mapped and visualized.

Mapping: Geographic Information Systems: Essentially the creation of 3D modeling within an computer environment of the ocean.

Satellites: Can detect and observe the ocean characteristics.

Semipermeable Membrane Devices: Used to collect various microbes for analysis of bacteria environment.

Sonar: Uses sound to detect area around a submersible, and well as figure out water depth.

Sonde and CTD: Collect data on a multitude of things, primarily temperature at different depths and conductivity of the water.

These sensors and technologies are what help make the data we can use to analyze and help predict our oceans health and patterns. All of these tools collect different data, from the directions of currents, temperatures, bacteria, as well as mapping data. With all of this information, it is now becoming a challenge to ask the right questions of it. That is the main issue in the field right now is about asking the right questions. With the oceans being the least explored and mysterious place on earth we are very unsure how the system works as a whole and the cause and effect major events have on it.

2 PROBLEMS THE OCEAN FACES

The ocean, is a big an complex ecosystem, no only that, but is rather sensitive to change. Given humanities treatment of the ocean and it inhabitant, it has taken quite a few damaging blows to say the least. As of now there are multiple threats to the ocean. There is the climbing temperatures, which effect the coral reefs and ocean level. The ever present amount of plastic in the ocean. From large vortex patches in the middle of the oceans that are miles wide, to the small micro plastics fish are eating, and intern are being eaten by us. Not to mention the catastrophic effects of the multiple oil spills, and bleaching incidents that are destroying the coral reef habitats. Which coral reefs have two fold purpose, one is the habitat, the second is the reef itself which helps remove a lot of carbon dioxide from the atmosphere. These are the most pressing issues as of right now but there are many more. With the use of big data and collection through the sensor above we can start learning how to take action and where.

2.1 Collection Efforts

As of right now the collection of data is not a collective effort globally. [2] However there are many organizations that are doing their best to collect and analyze the data. One of these organizations is The National Oceanic and Atmospheric Association. Which is collecting massive amounts of off coastal based laboratories. The use of data is to help get an unbiased amount of data to help predict the health of the ocean and its effect on climate change and rising sea levels. That way it can help coastal cities and states, plan for the future within the next couple decades.

Even with organizations like this around the globe, there is still not enough data to truly predict and know the health of our oceans. It will take a global effort and many more sensors, and a lot more interest from the public. This will help with funding for the scientist and the tools that will need to be developed and maintained for future exploration.

With enough data we could finally see how our planets oceans work. It could help choose the best route for boats, based off currents. It could help prevent major disasters by helping cities, states, countries be prepared for hurricanes weeks in advanced. It would allow us to see how events in different parts of the globe would have an effect elsewhere.

3 THE DATA

As with any situation where massive amounts of data are involved, things can get bit complicated. In the case of ocean data, you have data from multiple sources. Ranging from satellites, to buoys, to bacterial analysis. Not only that but the complexity with all of this data is that the ocean is one of the most dynamic things on earth. Unlike things on land, though always moving, are usually predictable. The ocean does not give any such luxury as it is all connected and always in motion. Given this fact, it is hard to collect data for the ocean as a whole. At least for right now. However, it is possible to get fairly reliable data for a set area.

3.1 Integration

Making sense of all this data is the main issue that most scientist face. So integration is a primary concern. When modeling the data especially there is a lot of work to make the data come together and makes sense. In one case for modeling salinity of different parts of coastal regions, they match up GPS data points and observations and time with the data from the senors that collect it. The matching of all this information allows you to integrate the data together for correct mapping of a given location. [3] This is just a small example of the amount of effort to pull all the data together to produce visualization of those results

4 SOLUTIONS

Though the efforts have not come together globally as of yet, plans to make that happen are in motion. Just this June, the United Nations met for a week long discussion on how to make our oceans healthier with emphasis on global cooperation. [4] Unlike the pairs agreements for global warming, this meeting will not come to a signed contract, rather an understanding that be every country will to be accountable for some part of the oceans health. And encourage international cooperation to help solve our biggest issues.

4.1 Hopeful Future

At this point there is no clear cut solution to solve all issues that face our oceans. And we may never collect all the data that we need to make the most informed decision. However with the data collected thus far is clear that we need to do. And though we do not have it all the answers, we have enough data to start asking the important questions to set us on the right track.

ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski and all the TA's for their help. As well as other students and their contribution to collective learning.

REFERENCES

[1] 2013. Observing Systems and Sensors. (2013). http://oceanexplorer.noaa.gov/technology/tools/tools.html

[2] Nishan Degnarain. 2017. How data can heal our oceans. Web Page. (August 2017). https://www.weforum.org/agenda/2017/08/how-data-can-heal-our-oceans/

[3] Yingjian Liu, Meng Qiu, Chao Liu, and Zhongwen Guo. 2016. *Big Data in Ocean Observation: Opportunities and Challenges.* Springer International Publishing, Cham, 212–222. https://doi.org/10.1007/978-3-319-42553-5_18

[4] Todd Woody. 2017. Crunch Time: What Big Data Reveals About Ocean Health. (July 2017). https://www.newsdeeply.com/oceans/articles/2017/07/06/crunch-time-what-big-data-reveals-about-ocean-health

Big Data Analytics for Municipal Waste Management

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

Mani Kumar Kagita
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 43017-6221
mkagita@iu.edu

ABSTRACT

Waste management is becoming a greater concern for cities and municipalities around the world because of the continual increase in population and waste. Big data analysis has the potential to not only help assess the current waste management strategies but also provide information that can be used to optimize the systems used in various institutions, local governments, and companies, among others.

KEYWORDS

Waste Management, Big Data, Local Government, HID305, HID319, i523

1 INTRODUCTION

In the current fast-paced society, production of goods continues to increase, and new distribution chains continuously change. The generation of waste and the deprecated goods -from now on referred to as solid waste- has increased over the past ten years, rising from approximately 0.64 kg per person per day of solid waste to approximately 1.2 kg per person per day. Projections estimate that this number is expected to increase to about 1.42 kg by 2025 [15]. This continual change and increase make waste management a more complex and more intensive endeavor.

While the quantity of waste itself can and should be reduced by conscious use and discipline in recycling and reusing items, local governments and waste management companies can also make modifications to their systems to reduce waste in the actual collection and disposal.

Because of this, different local governments and organizations have recognized the need to develop more elaborate regulations to control the different features, segments, processes involved in waste disposal from the moment the material is discarded by the consumer to the moment the material reaches its ultimate destination, such as a recycling plant or landfill. This set of systematic regulations is called solid waste management. Solid waste management has changed over time. What used to be systems designed and implemented based on local needs and convenient disposal as moved into extensively researched and implemented management systems that consider complex multivariable and dynamic sources of data [1].

Currently, data used in waste management is collected from many sources and varies depending on the types of solid waste and the rate of disposal of a particular population. Because of the diversity of data available (including types of waste, weight collected, the location of collection and disposal), the quality of the data must be continually monitored and assessed. [6]. New management systems that seek to optimize waste management must collect large volumes of data from various data sources on a daily basis in order to compile information essential for optimization. Multivariate data analysis methods provide exploratory data analysis, classification and parameter prediction using this data [4].

2 MUNICIPAL SOLID WASTE MANAGEMENT

There are various factors that contribute to the complexity of optimizing waste management systems and the types of data needed for analysis to do so. These include differences in composition of waste, environmental and logistic needs for individual communities, the changing forms of handling various materials such as recyclables, and the various ways that waste can be disposed of.

There are significant differences between the general composition of the waste generated in rural areas and the waste produced in urban areas. The waste produced in the later is profoundly influenced by the culture and the practices of our modern society. This distinction means that there will inevitably need to be differentiation in waste management practices and systems based on each communities needs, commerce, economy, and practices [6].

Municipal Solid Waste (MSW), commonly known as garbage or trash, consists of items that residential, commercial, institutional and industrial sites generate. This is much different from the type of waste produced, for example, in an agricultural community. Urban, or municipal, waste can be the result of everyday activities, such as leftover food, plastic bottles, packaging material, wooden furniture, electrical and electronic appliances, glass, medical waste, card boards, waste tires, office wastes, consumer goods, among others. Each type of waste needs to be handled in distinct manners.

The amounts of solid waste and its composition vary depending on the country, place, and activity performed at the site where the waste is generated [6]. For this reason, every process related to waste management (transportation, storing and final disposition, among others) must be engineered and tailored to fit the specific needs of community, organization, or local institution. The data collected for analysis will also be specific to the needs and realities of each entity, which is why waste management plans are not universal or easily applied from one community to another.

The forms of handling waste have also changed over the years, affecting the best ways to manage waste. One example of this is the increasing ability to separate and recycle various materials. Data on the recyclable material, dividing waste and recyclable material, how materials are sent or how they are disposed of has become significant. According to EPA statistics from 2014, Americans generated about 258 million tons of MSW of which more than 89 million tons is recycled and composted. This is an equivalent to a 34.6% recycling rate compared to 6.4% in 1960. With the increase in recycling capabilities and consciousness, adaptations must be made to the management systems [9].

Another variable that is how waste is disposed of. Americans have used the energy production process to combust approximately 33 million tons of waste, while as much as 136 million tons of waste ended up in landfills during the same year [9]. Local governments must determine which form of waste disposal is more efficient and cost-effective based on their unique context. They must take into consideration access to and distance from disposal facilities, types of waste being disposed of, and specific environmental implications of each process based on their geographic, geological, and environmental context.

In waste management, decision-makers are and will continue to be forced to make choices. When they develop or implement plans, they have a choice as to what information they will or will not consider for analysis. They have to choose what factors are influential in their jurisdiction. They make choices about routes, resources, and all the details between pick up to disposal. These choices can be classified as fortuitous, good, or optimal [1].

Fortuitous decision-making has no scientific base; the person who is in charge of making decisions must always try to solve the problem with little or no research or data. On the other hand, good decision-making is primarily based on experience, comparison of elements and trial and error. Optimal decision making, however, requires understanding and analysis of techniques and technologies provided by other fields [1].

This is where Big Data comes in.

3 BIG DATA AND WASTE MANAGEMENT

Big data has the capacity to facilitate analysis of information so that better decisions can be made.

In Big Data, the expert in charge takes many data sets coming from diverse and dynamic data sources and applies technologies to analyze these data sets. Authors of papers and books like "The Fourth Paradigm" state that big data exploration works to find patterns in data by analyzing the trends and outliers found in the data sets mentioned above, to generate knowledge [14].

Characteristics (such as the large volume of data generated in waste management, how dynamically the data is generated, and the variety of formats in which the data comes) make the task of producing knowledge an ideal task for Big Data. Scientists can interpret the findings of Big Data in a way that allows individuals and institutions involved in waste management to make optimal decisions [20].

4 SOLUTIONS FOR EFFECTIVE WASTE MANAGEMENT

4.1 Examples of Big data and waste management

Various institutions around the world have explored and implemented Big Data analytic to optimize their waste management systems based on their unique needs.

In Manchester, England, the Greater Manchester Waste Disposal Authority, England's most significant waste management institution, has started to use Big Data to better orchestrate the waste management services they provide. To get the most out of their Big Data approach, the Greater Manchester Waste Disposal Authority is working in collaboration with the research done by the University of Manchester. Together, they are trying to create environmentally sustainable solutions for Manchester, and are attempting to develop optimal solutions to the 1.1 million tons of waste that Manchester produces every year. While their work has not been implemented yet, they have recognized the need for both research and partnership with research institutions to compile and analyze data [19].

Another example of a local government implementing Big Data Solutions is the city of Songdo, South Korea. In said city, every citizen needs to use a chip card while disposing of their garbage. Data collected from these chip cards is being used to analyzing on the quantity of disposed waste and their locations. Each trash bin is incorporated with sensors to provide the height of the garbage accumulated in the reciprocal, temperature, and air pollution levels. These multiple parameters help municipal authorities forecast ideal times to collect the trash and optimize the routes to save time and expenses [19].

Researchers in Ethiopia are combining socioeconomic data alongside geographic data in order to get a clearer understanding of the patterns of how household waste is being collected and distributed. This study helped local authorities to better manage waste practices in urban areas [19].

A group of researchers from the University of Stockholm is using Big Data to identify how to optimize waste management routes in their city. By using a wide variety of data sets collected from various sources, roughly around half a million entries including trash bin locations, weights, and truck routes, researchers have developed waste generation maps of Stockholm. This research has helped reveal various inefficiencies in the current waste management system and will be integral in helping them improve their local waste management [19].

4.1.1 Vehicle Routing Problem. Vehicle route optimization is one of the primary concerns in waste management. It is termed as Vehicle Routing Problem (VRP) [7]. While routes may be designed based on geography, merely looking at a map and designing what seems to be functional movement through the city, this is underestimating the extent of the factors involved. There are multiple factors that either directly or indirectly affect waste collection that can also be analyzed and considered while designing routes. Common known factors that influence vehicle routing are the type of vehicle, vehicle capacity, number of collection stops, volume per capita and the route length.

Two of the most fundamental VRPs are the Travelling Salesman Problem (TSP) and the Chinese Postman Problem (CPP) [2]. However, when too many constraints and attributes are considered, both of the TSP and CPP become more difficult problems to solve. Many researchers have studied and published articles on waste management vehicle routing problem VRPs, and yet it is a persistent problem. Mathematical models need to be developed to provide city administrators with tools to make effective long-and short-term decisions relating to their municipal disposal system [3].

In addition to the typical known factors causing VRP problems, indirect attributes must also be taken into consideration. Time of day, traffic, weather patterns, energy prices, demand fluctuations, vehicle health, dump site inventory also have potential to affect route efficiency.

A research team at OSI came up with a better solution for solving VRP problems using Big Data technologies. Mixed Integer Programming (MIP) formulation was designed to interact with millions of attributes in live environments providing real-time decisions to optimize the VRP. Big Data technologies are being used to enable prediction of travel times and forecast and address demand on a tactical time line. This approach has helped improve VRP forecasting between 5% to 10% [12]. Any improvement, even less than 5% created on VRP is a significant improvement [13].

Scientists and engineers have designed a system to collect data that communicates between waste reciprocals, waste collection vehicles and a central system using sensors that measure how full containers are and can send real-time data [10].

The data collected can be used to optimize routes and space in the waste collection vehicles. Sensors can identify which containers do not need to be collected, allowing them to shorten routes when containers do not have sufficient volume to be collected. In addition, sensors can measure the remaining capacity of collection vehicles, allowing them to extend routes when they still have cargo space, ensuring that they return to the disposal center only when they are carrying a full load. After the implementation of this system, the city was able minimize inefficient travel and fewer collection vehicles were needed. It was estimated that in three years, the expense of purchasing and implementing this system would be recovered [18].

In addition to the real-time data collection that helps optimize collection and routes, remote self-diagnostics are also helping optimize vehicle use. By being able to monitor the vehicles use and health, maintenance and repair are more efficiently managed. Through this system, managers are made aware of parts that need to be ordered in advance, limiting the time that the vehicles may be out of service. Hand held devices have been developed for service verification, further helping minimize external providers or assessments, optimizing asset management and costs [11].

4.1.2 Problem of Landfill Disposal. A municipal solid waste landfill (MSWLF) is an individually isolated area of land or a trench where the household waste is collected and stored. These landfills are designed to store municipal solid waste as well as other wastes like construction and industrial waste. These landfills can be open-pits or below ground refuse chambers. In recent years, it's becoming more expensive to operate and maintain these landfills as well as to protect the environment liquid pollutants that drain from the waste and which can cause water and surface contamination. These liquid pollutants are commonly called as leachate. Leachate forms when water originating from rainfall or groundwater dissolves gradually through a porous surface and dissolves the chemicals from the refuse especially if the protective layer of the landfill allows liquid or gases to pass through it.

Leachate contamination problems have become more problematic in older landfill sites where they lacked appropriate barriers above or below the landfills. This contamination has been found to cause pollution which may be a cause of diseases affecting the citizens residing in close proximity to these landfills. In modern days, Governments across the world are motivated to prevent environmental contamination due to waste disposal and to reduce the size and expansion of landfill deposits. Steps are taken to reduce the number of landfill areas and to extend landfill capacity at current sites.

Landfill disposal, in itself, has its own set of complex factors including measurement and control of gases, management of waste water, and precipitation patterns affecting water volume and run off in the geological basin.

Identifying these factors creates an opportunity to develop sensors that could collect meaningful data that would be an asset to optimizing landfill disposal. Sensors could measure gas emission, water volume in disposed waste, and precipitation. Having this information would allow data scientists to develop systems that would identify when disposal would need to be varied, how water treatment plants could efficiently manage runoff and basin water and how to communicate with water treatment plants to use more or less of their mechanical resources based on real-time needs.

Manual calculation would be time-consuming subject to human error and physical observation opens space for inefficiency and delay. In order to compute these calculations promptly so that the information could contribute to optimization, it would be ideal to develop a computer program to compute the mathematical operations. These calculations could translated into real time information that would help manage the systems in place. This would contribute significantly to optimization of water treatment, for example, in landfill disposal [1]

Waste management is already in the process of transforming from using older methodologies to modern Big Data technologies. Big Data has a lot more potential to eradicate many of the problems faced by government in waste disposal. Most of the government organizations are taking a leap to minimize disposal of waste and to achieve "zero waste" goals [17] .

A program in government of District of Columbia called "Zero Waste DC" has initiated developments in order to provide resources that will help its residents move towards zero waste. Its primary motive is to divert 80% of MSW to recycling or source of energy where the remaining 20% non recoverable waste will be sent to landfills. By collecting vast amounts of data from all sources, they will analyze the date collected to implement cost-effective strategies for converting waste to re-usable resources, improving environmental conditions, taking measures on human health, reducing greenhouse gas emissions and conserving of natural resources. By obtaining data from the sensors measuring contamination of air with harmful gases, water purification levels from the nearby lakes, physical properties as well as chemical properties of the soil, Big Data will be able to help analyze this data and provide solutions to prevent further pollution and help design better systems in the future [17].

One of the reasons for Big Data to have such high volumes of information in these examples, is that the sensors that collect the data gather it on a daily basis and from different sources like digital meters, sensors and social media.

In some cases it is possible to read information from the sensors placed in recycle bins, the data they provide allows the company in charge to identify the type of waste is being disposed in the trash bins and which resident does it. If somebody places hazardous materials in the bins, authorities can detect them before being collected by the waste management vehicles. Besides preventing unwanted incidents, this kind of information can be used to develop tailored

waste management training for the public. Big Data Analysis can also be employed in other areas of education, to encourage citizens to use food scraps as fertilizer and even develop strategies like "farm to farm" practices where the public institution can promote urban organic farming [16].

On the same line, the use of Big Data technologies employed in the waste disposal process, provides visibility to the amounts and characteristics of the organic waste, particularly when it comes to food waste. Identifying inefficiencies in food management, can be used to plan and improve the food production and transportation chains. These and other applications of Big Data, enable food related companies to make effective decisions in their purchase and procurement departments, as well as their management of their own organic waste [5].

United States' Environmental Protection Agency (EPA) is using Big Data in their waste management and recycle research. Big Data Analysis is used to identify quantities of solid waste being recycled and informing the public about the benefits they may receive when they recycle. A Municipal Solid Waste Characterization Report issued by the EPA in 2007, contains the data collected on generation and disposal of waste in the United States within the 30 years prior to the release of said report. This data is being analyzed to measure waste reduction and plan recycling programs across the country [8].

4.2 Statistics and Waste Management

The data collected from the different activities related to Municipal Waste Management, can be run through mathematical models developed to predict behavior and understand causes to complex issues. In other words, these data and applying statistics can be used by Big Data professionals to draw conclusions and foresee possible outcomes as related to effective policy making and actual waste production, collection, and disposal patterns.

In the field of statistics and probability, there are many data analysis methods that are used to study waste management and production, but the two most popular are: PCA (Principal Component Analysis) and PLS1 (or Partial Least Squares Regression). The Principal Component Analysis reveals relevant parameters within a large parameter set. This allows the researcher to find the most significant and essential properties of a sample when studying a particular question. On the other hand, scientists use the Partial Least Squares Regression to identify in two matrices, the major internal and external correlations [4].

There are many applications tools and packages that are well known in statistics that can be applied to waste management, some of them are SPSS, Canoco, The Unscrambler and R [4]. There is also a mathematical modeling language developed particularly for formulating and solving optimization problems using linear programming, called Lingo. Lingo optimization software uses the branch and bound methods to solve problems similar to the ones found when studying waste management [1]

These resources are valuable to the continual analysis of data being collected, continually helping in waste management optimization.

4.3 GIS Analytics

When it comes to Geographical Information Systems (From now on GIS), there are multiple software and hardware options in the market. From paid software like ArcGIS to Open and free software like GVSIG, some solutions can help interpret large data sets, apply statistics and algorithms of different kinds and display them in a way that refers to a geographical space.

Two optimal routing algorithms have been used to calculate routes for waste collection; Solomon's insertion algorithm and a clustering algorithm. Data was used to help create more efficient routes by minimizing driving distance while taking into consideration factors like lunch breaks (time windows) that affected the distribution of human resources and time/vehicle management. By adding vehicle depots, by rerouting, and by sharing various routes, they predicted that they would be able to shorten vehicle expenditure by as much as 10,000 km.

An additional routing algorithm was used to present information on the environmental significance for optimizing waste collection routes. The study assessed the use of roll-on/roll-off containers while also considering schedules and lunch breaks. Through the utilization of an adaptive large neighborhood search algorithm, alongside a clustering method, a residential waste collection was analyzed, taking into consideration actual pick up points and employee lunch breaks. Using this information, time windows and starting conditions were adapted alongside the routes to reduce the total distance by as much as forty-five percent [18].

5 CONCLUSION

Waste management has been a growing concern and will continue to be an important area for optimization as both consumer waste and population increase. As institutions, governments, and individuals look to assess and optimize resources, minimize cost and create less of an environmental impact of waste management, Big Data has the potential to continue to help provide the information needed for future advancement.

There are various tools being used to optimize the different waste management practices, and there is space to develop additional tools to continue to the information available to decision makers.

One of the main reasons to use Big data in Municipal Waste Management is to provide local governments with tools that would facilitate the implementation of systems to more efficiently manage how much, where and the growth rate of the material that the community dispossess. Optimizing the waste management systems can help minimize the environmental impact of the community's actions, reduce pollution, increase rates of recycled materials, optimize routes to reduce time and expenses, among others. By using Big data, local governments can also track the number and quantity or weight of disposals at different locations; this information can be mapped to reveal the locations of the most significant waste generators. This information will help entities develop specific strategies to reduce waste and help implement permanent solutions for better environments.

Waste Management is not only government issue. Citizens should take the initiative and educate others on how to recycle and reduce waste. With the help of the collected data, governments

can notify the different entities (individuals, communities, companies and other organizations) a equip them through education and awareness, and also share valuable information about the importance of waste management through different media, such as mobile phones, email, among others.

Local governments have just started to adopt Big Data technologies for solving problems involved in MSW, but there is plenty of room for growth and further use and development. By using Big Data Analytic, large amounts of data sets pertaining to specific communities waste, routes, and disposal can be used to identify trends and patterns that could highlight opportunities for improvement. Big Data can play a significant role in managing cities more efficiently, benefiting not only those managing the systems, but the communities in which they are implemented as well.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Mohsen Akbarpour Shirazi, Reza Samieifard, Mohammad Ali Abduli, and Babak Omidvar. 2016. Mathematical modeling in municipal solid waste management: case study of Tehran. *Journal of Environmental Health Science and Engineering* 14, 1 (18 May 2016), 8. https://doi.org/10.1186/s40201-016-0250-2

[2] Jeroen Belin, Liesje De Boeck, and Jonas Van Ackere. 2012. Municipal Solid Waste Collection and Management Problems: A Literature Review. *HUB RESEARCH PAPERS 2011/34 ECONOMICS & MANAGEMENT* 48, 34 (11 2012), 1–5.

[3] V. N. Bhat. 1996. A model for the optimal allocation of trucks for solid waste management. *A model for the optimal allocation of trucks for solid waste management.* 14 (1996), 87–96.

[4] K. Bofkhm, E. Smidt, and J. Tintner. 2013. Application of Multivariate Data Analyses in Waste Management. In *Multivariate Analysis in Management, Engineering and the Sciences*, Leandro Valim de Freitas and Ana Paula Barbosa Rodrigues de Freitas (Eds.). InTech, Rijeka, Chapter 02, 15–16. https://doi.org/10.5772/53975

[5] Frank E. Celli. 2016. Data analytics: The most effective approach for a zero waste solution. WasteDive. (09 2016). http://www.wastedive.com/news/data-analytics-the-most-effective-approach-for-a-zero-waste-solution/425424/

[6] R. Chandrappa and J. Brown. 2012. *Solid Waste Management: Principles and Practice.* Springer Berlin Heidelberg, Berlin. 47–63 pages. https://books.google.com/books?id=kUOwuAAACAAJ

[7] G. B. Dantzig and J. H. Ramser. 1959. The Truck Dispatching Problem. *Management Science* 6, 1 (10 1959), 80–91.

[8] EPA. 2007. METHODOLOGY FOR ESTIMATING MUNICIPAL SOLID WASTE RECYCLING BENEFITS. epa.gov. (11 2007). https://www.epa.gov/sites/production/files/2015-09/documents/06benefits.pdf

[9] EPA. 2014. Advancing Sustainable Materials Management: Facts and Figures. U.S. Environmental Protection Agency. (2014). https://www.epa.gov/smm/advancing-sustainable-materials-management-facts-and-figures#Materials

[10] Maurizio Faccio, Alessandro Persona, and Giorgia Zanin. 2011. Waste collection multi objective model with real time traceability data. *Waste management (New York, N.Y.)* 31, 12 (08 2011), 2391–405. https://www.ncbi.nlm.nih.gov/pubmed/21821406

[11] Megan Greenwalt. 2017. What the Growth of Big Data Means for Waste & Recycling. FLEETS & TECHNOLOGY. (03 2017). http://www.waste360.com/fleets-technology/what-growth-big-data-means-waste-recycling

[12] Vijay Hanagandi. 2013. A New Paradigm to Solving Vehicle Routing Problems. (09 2013). https://osiblogdotcom.wordpress.com/2013/09/23/a-new-paradigm-to-solving-vehicle-routing-problems/

[13] Geir Hasle, Knut-Andreas Lie, and Ewald Quak. 2007. *Geometric modelling, numerical simulation, and optimization: Applied mathematics at SINTEF.* Springer, Berlin, Heidelberg, Oslo,Norway.

[14] 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

[15] Perinaz Hoornweg, Daniel; Bhada-Tata. 2012. *A Global Review of Solid Waste Management.* Number 15 in Urban Development Series. World Bank, Washington, DC, Urban Development & Local Government Unit World Bank 1818 H Street, NW Washington, DC 20433 USA. https://openknowledge.worldbank.org/handle/10986/17388

[16] James Kobielus. 2012. Reuse, Recycle, Compost: A New Level of Insight into Garbage. Robert Reed, Evangelist at Recology, zerowasteIBM. (06 2012). http://www.ibmbigdatahub.com/blog/reuse-recycle-compost-new-level-insight-garbage

[17] Cole Rosengren. 2017. San Francisco expands recycling list, shrinks refuse carts on 'zero waste' crusade. WasteDIVE. (10 2017). http://www.wastedive.com/news/san-francisco-expands-recycling-list-shrinks-refuse-carts-on-zero-waste/506700/

[18] Hossein Shahrokni, Bram Van der Heijde, David Lazarevic, and Nils Brandt. 2014. Big data GIS analytics towards efficient waste management in Stockholm. In *Proceedings of the 2014 conference ICT for Sustainability.* Atlantis Press, Proceedings of the 2014 conference ICT for Sustainability, Department of Sustainable Development, Environmental Science and Engineering, Industrial EcologyRoyal Institute of TechnologyStockholm, Sweden, 140–147.

[19] Mark van Rijmenam. 2016. *How Big Data Shapes Urban Waste Management Services in Manchester.* techreport. University of Technology, Sydney. https://datafloq.com/read/how-big-data-shapes-urban-waste-management-service/662

[20] 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 in general: Mani Kagita, Andres Castro Benavides: work equally split between.

Editing: Andres Castro.

[6]

Automated Diagnostic Code Extraction in Electronic Medical Records

Nicholas J Hotz
Indiana University
nhotz@iu.edu

ABSTRACT

Electronic medical records (EMRs) play an increasingly important role in healthcare. However, the rapidly growing volume of text in EMRs creates challenges for information extraction (IE). As such, many research institutions are developing computer-based systems to automate EMR structured IE. This paper investigates the processes, the challenges, and the current state of automated IE of EMRs with a specific focus on automated systems that extract ICD9 codes from clinical text. While automated system performance has caught up to the accuracy of manual coding under specific circumstances, automated code extraction remains mostly an academic exercise. To extract value from their work, researchers should shift their focus away from highly specialized algorithms that work in isolation and instead collaborate with industry to develop augmented intelligence systems that help make coding professionals more effective.

KEYWORDS

i523, HID210, Natural Language Processing, Medical NLP, Clinical NLP, Information Extraction, Clinical Coding, Healthcare Big Data

1 INTRODUCTION

Demand for structured health data continues to grow [20], and the adoption of electronic health records (EMRs) generates new opportunities to improve clinical care, administrative processes, clinical workflows, and patient outcomes through higher quality, more accurate, more consistent, and more easily accessible documentation [14] [17].

However, the size, growth, and textual nature of EMRs render traditional software and hardware unable to effectively manage healthcare big data [18]. Healthcare data in the United States reached 150 exabytes in 2011 with Kaiser Permanente, Califronia's health network, reportedly having between 26.5 and 44 petabytes alone [5]. The volume of healthcare data is doubling every 12-14 months [7], and the diversity of this data further complicates its analysis [10]. Much of it is stored in narrative form which describe patients, their own and their family's medical history, their personal lifestyle, and their current medical conditions [14]. Although convenient for documentation, narrative text is difficult for computer systems to interpret as coded data that can support research, provide clinical knowledge and performance information, and improve patient outcomes [14] [20].

Commonly studied clinical NLP problems include de-identification [23], the development of patient problem summaries [8], and diagnostic code extraction [15]. This paper focuses on diagnostic code extraction which is the process of converting EMR clinical narratives into appropriate medical codes such as ICD9 (the standard medical diagnostic hierarchical taxonomy system in the United States until September 30, 2015). Perotte et al. describe both the ICD9 and the more recently adopted ICD10 taxonomies as "organized in a rooted tree structure, with edges representing is-a relationships between parents and children" [15]. Kavauluru et al. explain that the ICD9 and ICD10 leaf nodes are codes that provide specific information used for "billing and reimbursement, quality control, epidemiological studies, and cohort identification for clinical trials" [12].

Currently, coding professionals and clinicians manually extract diagnostic codes from EMRs which is expensive, inefficient, and has become increasingly complex due to various factors including the expansion of payment systems, new reporting requirements, increased oversight and regulation, and the increased volume of EMR data [3] [17] [20] [23]. This complexity limits manual coding accuracy. Manual coders often disagree [16] and are more specific than sensitive in their code assignments [2]. Errors are prevalent; for example, a Swedish study of 4,200 patient records found errors in 20% of the main diagnoses [23]. Over-coding can lead to fraud if healthcare providers bill for services not rendered while under-coding prevents providers from earning reimbursements for valid conditions and services [15].

Since the 1990s, researchers have tried to improve the coding processes through automated coding and classification technologies [11]. Stanfill et al. in their comprehensive literature review in 2010 describe these automated coding systems as "a variety of computer-based approaches that transform narrative text in clinical records into structured text, which may include assignment of codes from standard terminologies, without human interaction." They cite that the American Health Information Management Association asserted in 2004 that, "The industry needs automated solutions to allow the coding process to become more productive, efficient, accurate, and consistent." Yet, Stanfill et al. conclude that the relative performance of automated systems to manual coding is not yet known [20]. As of 2008 and still in 2015, automated systems are still mostly used for research purposes with few applications in use by practitioners [14] [23].

2 EMR INFORMATION EXTRACTION CHALLENGES

Several challenges have slowed the development of clinical text NLP applications, which lag behind NLP applications in other fields [4]. Meystre, et al. attribute the lack of shareable clinical data as the biggest challenge [14]. Large annotated corpora are needed to develop effective machine learning algorithms that can classify roughly 17,000 possible ICD-9 codes and 68,000 ICD-10 codes whose frequency distributions are highly skewed [1]. However, clinical information needs to be de-identified (which itself is a challenging problem) in order to comply with privacy concerns and regulations

such the USA's Health Insurance Portability and Protection Act (HIPAA) and the European Union's General Data Protection Regulation (GDPR); as a consequence, large corpora typically remain siloed within individual healthcare systems and are rarely available for outsiders [14] [20].

As a related problem, even when corpora are available, the annotation process is time-consuming, expensive, and traditionally relies on domain experts and linguists [14] [23]. Given the highly specific sublanguages of clinical text, general NLP systems perform poorly on cross-domain clinical texts without these comprehensive annotated corpora. Consequently, much of the development in clinical text NLP occur in siloes and is not used outside of the laboratory in which they were developed [4].

In addition to the lack of shared annotated corpora, Meystre et al. present four challenges that hinder the development of effective clinical text IE. First, clinical narratives contain ungrammatical phrases with short-hand abbreviations and acronyms. About a third of these short-hand texts are overloaded (a single unit may have multiple meanings) which can be challenging for human interpretation and even more challenging for computer interpretation. Second, the rate of misspellings, around 10% [19], is higher than most texts complicates many NLP techniques. Third, clinical texts often contain long series of non-text information, such as laboratory test results, which makes sentence segmentation difficult. Forth, institution-specific pre-formatted templates that appear in clinical texts are difficult for interpretation and their meanings do not transfer to other institutions' information [14]. Chapman et al. discuss additional challenges including the inadequacy of de-identification algorithms, the lack of focus for NLP in non-English clinical texts, and the absence of common clinical standards [4].

Fortunately, recent progress is promising as explained in literature reviews by Delanis et al. (2014) and Velupillai et al. (2015). These publications praise the clinical NLP community for overcoming many of these hurdles by providing more annotated corpora, developing more advanced NLP tools specific to clinical text, leveraging partially-automated processes to facilitate the annotation of corpora, and focusing on multiple languages [6] [23].

3 EMR PRE-PROCESSING

To convert text to medical codes, clinical text flows through various pre-processing and context feature detection techniques. General pre-processing NLP tools are being adopted and specialized for medical texts including:

- **Language detection:** Multi-lingual studies may start with language detection algorithms, although some might still rely on manual detection [8].
- **Spell checking:** Clinical NLP spell checking uses standard dictionaries and medical-specific tools such as unified medical language system (UMLS) and WordNet [14].
- **Word sense disambiguation:** WSD allows the system to identify the correct meaning of a word that has multiple definitions; however this process is not as accurate with clinical texts as with general English (about 90% for general English and 80% for clinical text) [14].
- **Tokenization and sentence-splitting:** Tokenization splits text into smaller components called tokens which include words, phrases, and symbols [8] [21].
- **Part-of-speech tagging:** Also known as lexical analysis, POS tagging identifies a word's part of speech and its relationship with other words in a sentence [8] [14].
- **Parsers:** Parsers identify the sentence syntax, word dependencies, and expressions of interest [8] [14].

Context feature detection and analysis happen concurrently or following the above steps and identify how words and concepts are used in the context of the sentence. Clinical NLP systems often use a set of regular expressions and algorithms such as NegEx, NegExpander, TimeText, and ConText to define feature context. Notable contexts are negation (e.g. patient *does not* have a condition), speculative (e.g. patient *might have* a condition) temporality (e.g. to identify if the patient *has* or *had* a condition), subject identification (e.g. to identify if the condition belongs to the patient or someone else such as a family member), and severity (such as mild, moderate, or severe conditions) [14] [23].

4 REVIEW OF AUTOMATED ICD9 CODE EXTRACTION EFFECTIVENESS

To evaluate the effectiveness of automated systems, studies compare evaluation metrics against standards. Per Stanfill et al.'s literature review of 113 studies, 43% of studies use the gold standard comparison which uses at least two independent reviewers and an adjudication process to resolve inconsistencies, and 51% use the regular practice standard of one reviewer [20]. Although considered more reliable, gold standards are still prone to error [15]. The most commonly reported metrics include recall or sensitivity (69%), PPV or precision (46%), specificity (43%), and accuracy (25%) [20].

Most studies focus only on a specific subset of clinical texts or diagnoses such as subdomains like radiology [17], for specific diagnoses like congestive heart failure [9] or cancer [13], or to extract only attributes of patients like smoking status [22]. Although many of these studies achieve accuracy metrics comparable or even exceeding gold standards, their results are not generalizable for more comprehensive or practical purposes in the field [20].

However, two recent studies attempt to comprehensively extract ICD9 codes from large EMR sets. In 2013, Perotte et al. attempted to extract ICD9 codes from the clinical text of Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II), a publicly available database containing de-identified records of 40,000 ICU hospital admissions. They split the 22,815 discharge summaries, which contain 215,826 ICD9 codes (5030 distinct) into 20,533 training documents and 2,282 testing documents. Using a hierarchy support vector machine (SVM) classifier, Perotte et al. report an F-measure of 39.5% with a 30.0% recall and 57.7% precision. They also attempted a flat SVM which returned a 27.6% F-measure with 16.4% recall but with a higher precision (86.7%) [15].

Similarly, in 2015 Kavuluru et al. developed automated coding systems with 71,463 in-patient EMRs from the University of Kentucky Medical Center. They conclude that the best-performing automated coding method depends on the size and characteristics of the dataset. For smaller narratives in subdomains such as radiology or pathology, chain classifiers perform best because codes are

similar to each other. However, feature and data selection methods perform best with more comprehensive in-patient EMRs. Meanwhile, "for large EMR datasets, the binary relevance approach with learning-to-rank based code reranking offers the best performance." They reported a micro F score of 0.48 with codes that occur at last 50 times and a score of 0.54 for codes that occur in at least 1% of records [12].

5 CONCLUSION

Researchers are increasingly studying clinical NLP and diagnostic code extraction. However, the output of most research is limited to specific circumstances and has not yet been applied to practical use cases that improve the accuracy and efficiency of medical coding processes. Rather, the research community seems to evaluate its work in terms of algorithm accuracy metrics in their specific strength zones relative to the performance of human coders. Cross-domain medical coding studies are a step in the right direction toward a more practical approach which begins to mimic the reality faced by human coders.

However, the clinical NLP researchers should take this progress further, and collaborate with software engineers, HCI design specialists, business analysts, medical coders, and clinicians to develop practical augmented intelligence systems. These systems, which can include semi-automated recommendation and auditing support software solutions, can aid medical coding professionals in actual workflows to extract diagnostic codes from medical text. A workflow that leverages the strengths of algorithmic systems to shore up areas of human coder weaknesses can optimize medical coding efficiency and accuracy.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Stefan BERNDORFER and Aron Henriksson. 2017. Automated Diagnosis Coding with Combined Text Representations. *Informatics for Health: Connected Citizen-Led Wellness and Population Health* 235 (2017), 201.

[2] Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. 2005. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical care* 43, 5 (2005), 480–485.

[3] J Bishop, J Bronnert, J Cook, L Macksoud, C Mastronardi, M Morsch, K Phibbs, R Scichilone, and K Thibault. 2010. Automated Coding Workflow and CAC Practice Guidance. *Journal of AHIMA* 81, 7 (2010), 51. http://bok.ahima.org/PB/CACGuidance#.WchAZMiGOUl

[4] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. (2011).

[5] Mike Cottle, Waco Hoover, Shadaab Kanwal, Marty Kohn, Trevor Strome, and N Treister. 2013. *Transforming Health Care Through Big Data Strategies for leveraging big data in the health care industry.* Technical Report. Institute for Health Technology Transformation. 1–24 pages.

[6] Hercules Dalianis, Aurélie Névéol, Guergana Savova, and Pierre Zweigenbaum. 2014. Didactic Panel: clinical Natural Language Processing in Languages Other Than English. In *AMIA Annual Symposium 2014*. American Medical Informatics Association, American Medical Informatics Association, Stockholm, Sweden, S–84.

[7] Ivo D Dinov. 2016. *Volume and value of big healthcare data.* Technical Report. Health and Human Services.

[8] Crescenzo Diomaiuta, Maria Mercorella, Mario Ciampi, and Giuseppe De Pietro. 2017. A novel system for the automatic extraction of a patient problem summary. In *Computers and Communications (ISCC), 2017 IEEE Symposium on*. IEEE, IEEE, Heraklion, Greece, 182–186.

[9] Jeff Friedlin and Clement J McDonald. 2006. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *American Medical Informatics Association* 2006 (2006), 269.

[10] Sullivan Frost. 2015. Drowning in big data? reducing information technology complexities and costs for healthcare organizations. (2015).

[11] Ramakanth Kavuluru, Sifei Han, and Daniel Harris. 2013. *Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques.* Technical Report. Health and Human Services. 77–88 pages.

[12] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine* 65, 2 (2015), 155–166.

[13] Burke W Mamlin, Daniel T Heinze, and Clement J McDonald. 2003. Automated extraction and normalization of findings from cancer-related free-text radiology reports. In *AMIA Annual Symposium Proceedings*, Vol. 2003. American Medical Informatics Association, American Medical Informatics Association, Washington, DC, 420.

[14] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 35, 128 (2008), 44.

[15] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2013. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (2013), 231–237.

[16] John P Pestian, Christopher Brew, Pawe l Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and W lodzis law Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 97–104.

[17] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology* 279, 2 (2016), 329–343.

[18] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems* 2, 1 (2014), 3.

[19] Patrick Ruch, Robert Baud, and Antoine Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine* 29, 1 (2003), 169–184.

[20] Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association* 17, 6 (2010), 646–651.

[21] Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. Jena University Language & Information Engineering (JULIE) Lab, Jena, Germany, 49–57.

[22] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association* 15, 1 (2008), 14–24.

[23] Sumithra Velupillai, D Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. 2015. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics* 10, 1 (2015), 183.

What is the Role of Big Data in Health

Matthew Durbin, MD FAAP
Indiana University School of Medicine Department of Pediatrics, Division of NeonatologyfiRiley Hospital for Children
699 Riley Hospital Drive
Indianapolis, Indiana 46202fi
mddurbin@iu.edu

ABSTRACT

The United States healthcare system is in crisis. The cost is far higher than all other nations, and is exponentially increasing. In addition, for all of this spendig, measured health outcomes are far lower than most other nations. Big Data driven technology solutions may help solve the healthcare crisis.

KEYWORDS

i523, hid311

1 WHAT IS THE ROLE OF BIG DATA IN HEALTH

The current state of healthcare system in the United States is often described as a crisis. The term comes with good reason, as spending accounts for 17-18% of GDP, dwarfing other nations, and is exponentially rising at an unsustainable rate. For all of our spending, we have poorer health than most developed and many developing nations. The healthcare industry is behind in technology, with recent adoption of an electronic medical record, and prior reliance on paper charting. Communication is most often by decades old technology including phone or fax. Internet communication between healthcare providers, and with patients, is a recent novelty. We have to poorest health, including obesity due to poor diet, lack of exercise, and substance abuse. We pay more for pharmaceuticals than any other country, and most pharmaceutical budget goes to marketing as opposed to research and development. Meanwhile the business world is far ahead of healthcare.

Big Data has major potential to impact health. Massive data sets related to human health are compiled by insurance companies, pharmaceutical companies, public health institutions and research institutions. Big data will soon have a huge impact on improving the health, but there is a long road ahead. Much of the lag is due to serious issues with privacy and security. The healthcare industry should be able to overcome these obstacles as online banking and financial institutions have done. There is amazing potential with big data and healthcare, but a long way to travel. [4] Healthcare is making strides and big data collection is visible everywhere. The electronic medical record EMR is close to universal and is improving constantly. Medical resources are accessible around the world through smartphones, making medical libraries obsolete. Next generation sequencing technologies are able to measure the genetic contributions to disease that previously a mystery. Wearable technology and fitness tracking apps, nutrition apps are improving personal.

2 COST OF HEALTHCARE

2.1 The Current State

One of the most troubling issues facing the United States, and the world, is the increasing cost of healthcare. The problems are different around the globe. Much of the developing world lacks access to adequate healthcare, which is a serious problem. This paper focuses on a different problem, in the crisis facing the United States. Current healthcare spending is greater than 3 trillion dollars [3]. This makes up 17 percent of GDP. This number grows every year and is unsustainable. This number affects citizens deeply, and currently healthcare costs are responsible for 50% of bankruptcy claims in the United States [4]. All of this extra spending does not equal better health. In most measures of health, from infant mortality to life expectancy, the United States find itself far from the top. There are major issues at play ranging from a massive bureaucracy, to the poor health and obesity of participants.

2.2 The Future

It is projected that the average family will spend over 25% of income on to healthcare [4]. The problem is not projected to improve. As the *baby-boomers* age, the population over 60 with high cost chronic healthcare problems, increases exponentially. In Medical School, we were taught about this *silver tsunami* approaching the US healthcare system (prompting me to go into Pediatrics.) Many individuals, including myself, look to Big Data to uncover these problems and help fix them. Before it is too late. There are technology solutions including the electronic health record, medical reference technology, genomic medicine, telemedicine, wearable health technology, and personalized medicine.

3 ELECTRONIC HEALTH RECORD

3.1 Adoption of and EMR

Throughout history, medical records were taken on paper, but after 2000 the slow transition to electronic records began [7]. The handwritten records were kept in large file cabinets, and when records needed to be shared between physicians or institutions (across the country or across the street), the paper records were faxed over a telephone line. This technology is decades old. As technology raced forward with supercomputers and the worldwide web, medicine continued to use these antiquated forms of communication. Finally, government mandating forced healthcare systems into the modern era and electronic records went online. Currently over 84% of health records are online [4].

3.2 The Current State

A majority of healthcare systems around the world are under a government regulated socialized medical system which comes with a universal health record. The healthcare system in the United states is privatized, therefore the transition to EHR came with individual health entities purchasing a multitude of different EHRs. The problem comes in that a patient presenting to two different healthcare facilities, even if across the street or within the same building, will have two different medical charts that do no communicate with one another. The other problem comes with accessing this information. The two largest companies Epic and Cerner have a commercial interest, with a primary goal to increase revenue to the shareholder. It is exceedingly difficult for the nonprofit entities including academic centers and hospitals to access the patient information within the EHR. There is tremendous potential within the EHR. Beyond data collection, storage, data retrieval, and analysis, we should move towards real time guidance and guidelines for medical decision making to improve health.

4 KNOWLEDGE

Only 10-20 years ago, Hospital libraries and medical school libraries were once filled with books and journal articles. If a healthcare practitioner wanted information relevant to clinical care, they went to libraries to pour through the resources with exhaustive efforts. Today, those libraries are mostly void of books. Almost every individual in western medicine has access to a computer, and usually to a handheld device, capable of accessing far more information than could ever be stored in a library. There are massive information sources, such as PubMed, a gigantic repository of journal articles and books that is constantly being updated with new information. And Up To Date, a point of care medical reference commonly used on a handheld deceive, with evidence based clinical guidelines contributed by over 5,000 physicians [9]. The massive amount of data now accessible to most healthcare providers and scientists is changing healthcare rapidly. Still, there is much room for improvement as care is commonly delivered based on anecdotal evidence, and cost and quality should continue to improve.

5 NEXT GENERATION SEQUENCING

5.1 The Human Genome

The first human genome was sequenced in 2003[2]. This colossal global effort took over 10 years and thousands of scientists working at great expense. In the end, a private and public group collectively sequenced the first genome. Initially, the technology was extremely expensive and took great deal of time. Through technological advancements including sequencing cores and big data, the cost of the genome has plummeted. The 1000-dollar genome project is an attempt to make sequencing more affordable [4]. We are a long way away form being able to utilize the genome to deliver care. Bioinformatics expertise has lagged behind technology. Groups still do not agree on a standard way to process the information. Still this technology improves papidly, and recently a group published 24-hour genome sequencing for intended us in clinical decision making. Soon it may be a reality for physicians to utilize genomic information, whether about drug susceptibility, or prognosis, to guide medical care.

5.2 Beyond DNA

Initial estimates placed the number of genes at ¿100,000 [1]. Looking at the massive amount of diversity and the billions of unique human beings on this earth, this was a appropriate estimate. The current number is estimated somewhere around 20,000. The question is what accounts for the rest of phenotypic diversity and disease. The human genome project utilized whole exome sequencing. Whole exome sequencing involves sequencing the entire coding region, or exome, of the genome. This consists of around 20,000 genes and over 30 million nucleotides. The exome, though massive, consists of only 1% of the total genomic DNA. Many genetic diseases involve alteration of this coding exome but we are discovering that many diseases are due to problems outside of this coding region. Sequencing only 1% of the genomic material is a fraction of the time, cost, and burden of analysis, compared with whole genome sequencing, but we must move towards whole genome sequencing to capture all disease states. We have also come to realize that splicing and other post transactional regulation introduces much diversity. We have the technology to sequence the entire RNA transcriptome and the proteome as well. This produces a data set which dwarfs the genome and genomic DNA sequence information. These technologies are currently only utilized in the research setting. Despite our advanced technology, we have very little idea of how to interpret the data in a clinical setting. Again the bioinformatics expertise lags behind. There is amazing potential to advance knowledge and study human disease and a tremendous amount of big data analytics along the way.

6 WEARABLE TECHNOLOGY, NURITION AND WELLNESS APPS

Massive data sets exist, collected by insurance companies, in electronic health records, by pharmaceutical companies and by research institutions. There is another very exciting source of big data on the horizon, in personal wearable technologies, and also fitness, wellness and nutrition apps [4]. Individuals wearing FitBits, with fitness apps on their mobile devices, wearing smartwatches, etc. can track health and wellness measures in ways that once required inpatient hospital monitoring and sophisticated research lab settings. They track sleep and activity throughout the day and night. In addition, there are countless apps which track nutrition and health. People log meals and nutrition to keep accountable. Often these apps work with time tested and well researched diets including weight watchers, etc. This technology has already changed the way many individuals look at health and wellness. This exciting new dataset has great potential to advance human health and improve disease that may be the root cause of our healthcare epidemic.

7 TELEMEDICINE

Telemedicine involves a virtual visit between a physician and patient [6]. There are obvious benefits, especially when a patient population is spread across a wide geographic space either due to a high level of physician specialization, or a rural patient population. Highly specialized, but critical subspecialists are often in great shortage. This places a great burden on the available providers, with often unsustainable schedules. Video technology allows doctors, nurses and practitioners to visualize patients, perform a limited

physical, and to communicate with individuals at a distance. There is great potential to improve cost and reduce burden. There are limitations. Many physician specialists are values for their technical, hands on skills. Telemedicine is not much of a help, the technical procedures, such as inserting airways into the trachea of small babies, and insert central arterial lines into major vessels to deliver lifesaving medications, require hands on skills. The same goes for surgeons and other highly skilled technical professions. Interventional techniques and robotics are increasingly being used to perform procedures, but while these operations are performed, a surgeon needs to very close, in case unforeseen accidents problems necessitate a conventional correction. Procedural specialties are the greatest expense to our healthcare system and their procedural skills are a long way from being performed through telemedicine or robotics.

8 SOCIAL MEDIA

One interesting trend is the multitude of health information shared over social media networks. Blogs, columns, and posts providing information about nutrition and wellness, news stories, and information sharing. The story reporting googlefis flu prediction trends ahead of the CDC, based on search history, spread virally over facebook [5]. The field will continue to expand. wonderfully

9 PERSONALIZED MEDICINE

Wikipedia summarized personalized medicine as: “a medical procedurethat separates patients into different groupsfi!?with medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease.” [8] In a way the culmination of big data and health is with personalized medicine. In a hopefully not so distant future the electronic health record, pharmaceutical data and genomic data will provide a more tailored, affordable, and high-quality approach to healthcare. Hopefully healthcare will catch up with financial and ecommerce and in their ability to harness big data for good.

10 CONCLUSION

This paper highlights just a handful of technology driven big data solutions to our healthcare crisis. As Congress debates legislature to face this crisis, big data more harmoniously moves towards solutions. Better health without economic ruin is a reality and big data will play a major role. Much work is left to be done

ACKNOWLEDGMENTS

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

REFERENCES

[1] [n. d.]. ([n. d.]). Vanderbilt University: Introduction to Bioinformatics Course Lectures.
[2] Francis S Collins, Michael Morgan, and Aristides Patrinos. 2003. The Human Genome Project: lessons from large-scale biology. *Science* 300, 5617 (2003), 286–290.
[3] Centers for Medicare & Medicaid Services et al. 2014. National health expenditures 2012 highlights. *Online verfügbar unter http://www.cms. gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/National-HealthExpendData/Downloads/highlights. pdf* (2014).
[4] Geoffrey Fox. [n. d.]. Unit 6 Lectures. ([n. d.]).
[5] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.
[6] Maria Hernandez, Nayla Hojman, Candace Sadorra, Madan Dharmar, Thomas S Nesbitt, Rebecca Litman, and James P Marcin. 2016. Pediatric critical care telemedicine program: A single institution review. *Telemedicine and e-Health* 22, 1 (2016), 51–55.
[7] Erik WJ Kokkonen, Scott A Davis, Hsien-Chang Lin, Tushar S Dabade, Steven R Feldman, and Alan B Fleischer. 2013. Use of electronic medical records differs by specialty and office settings. *Journal of the American Medical Informatics Association* 20, e1 (2013), e33–e38.
[8] Wikipedia. [n. d.]. Personalized Medicine. ([n. d.]). https://en.wikipedia.org/wiki/Personalized_medicine
[9] Wikipedia. [n. d.]. UpToDate. ([n. d.]). https://en.wikipedia.org/wiki/UpToDate Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 2 Sept. 2016.

An Overview of Big Data Applications in Mental Health Treatment

Neil Eliason
Indiana University Online
Anderson, Indiana 46012
nreliaso@iu.edu

ABSTRACT

Mental health treatment presents with complex informational challenges, which could be effectively tackled with big data techniques. However, as researchers and treatment providers explore these applications, they find a lack of infrastructure and ethical concerns hamper their progress. A unified approach of developing an ethically informed data infrastructure is necessary to proceed.

KEYWORDS

i523, HID 312, Mental Health Treatment, Big Data, Data Infrastructure, Data Ethics

1 INTRODUCTION

1.1 Big Data

There is no immutable or standardized definition of big data. However, most conceptualizations include data with high volume (amount of data stored), velocity (frequency of data input or update), and/or variety (number of data sources or types), known as the "three v's". As these factors increase, they reach the so called "three v tipping point", where traditional methods of analysis do not meet operational needs. Here, big data analytic techniques are utilized to make these unruly collections of data useful. For example, text mining, audio analytics, video analytics, and social media analytics are specific techniques used to make low value data more organized, condensed, and useful. Then predictive analytics take this processed data, and create data models which can predict future outcomes. These can be divided into regression techniques, which identify ways groups rely on each other, and machine learning techniques, which look for patterns in validated test data and then apply them to an unvalidated sample [3].

1.2 Mental Health Treatment

Mental health difficulties are a common problem across the United States, and worldwide. Mental illness of some kind was prevalent among 17.9 % of Americans in 2015, and of that number 4% experienced serious functional impairment as a result [14]. A 2014 meta-analysis study estimated that the worldwide prevalence of mental illness was 17.6% and that 29.2% of people would be diagnosed with a mental illness over the course of their life [18]. The effects of these disorders on individuals and societies is costly. The US Center for Disease Control and Prevention estimated that 36,035 people died during a suicide attempt in 2008, and that 666,000 sought emergency room care for self harming behavior [2]. In 2013, the Social Security Administration reported that 1,947,775 persons received social security/disability benefits for either a mood or psychotic disorder, which is around 19% of all recipients [16]. It is estimated that mental health issues had a $100 billion cost on the US economy in 2002 [14], and in 2015 there were over 12,000 mental health treatment facilities in the US [20].

Mental health treatment attempts to address these pervasive and complex problems at an individual level. While this by nature results in a system that is heterogeneous and complex, treatment still follows a fairly consistent pattern. First the mental health issue is identified [1], then treatment interventions are assigned [21], and finally treatment progress is monitored [4].

The identification process involves mental health screening and assessment. Screening attempts to identify a person's primary mental health risks and needs for the purpose of directing them to appropriate sources. They tend to be narrow in focus and brief, which allows them to be easily disseminated to help filter people to the right level of care. Similar to screening, assessment aims to identify a person's mental health dysfunction, but does so in more clinically robust categories, typically resulting in a diagnosis [1]. Once a person's mental health issues have been clinically identified, then interventions are assigned. Those traditionally take the form of talk-therapy to develop effective change strategies, medication to reduce symptoms of mental illness, and supportive services such as case management to help coordinate efforts towards the person's goals [21]. Treatment monitoring is essential to the treatment lifecycle, as this is where clincians receive feedback regarding the effectiveness of the chosen interventions. While it is natural for clinicians to do this informally, more intentional methods are often overlooked [4]. This process requires an extensive data gathering effort, which traditionally is labor intensive and requires a large team of clinicians.

1.3 Thesis

There are large numbers of people struggling with mental illness, and their treatment requires large amounts of frequent data from various sources. This process as traditionally done is inefficient and labor intensive. Big data analytic techniques are designed to target this kind of data, and could greatly increase treatment effectiveness and scope.

2 BIG DATA APPLICATIONS IN MENTAL HEALTH TREATMENT

2.1 Screening and Diagnosis

Mental health screening is the first chance to direct people in the appropriate direction to meet their mental health needs. Methods that can screen larger amounts of people effectively are critical, as many people with mental illness are not connected with treatment. Several studies explored using social media to identify mental illness in the general population, and demonstrated potential to identify

issues on a large scale. Many attempted to identify depression by analyzing the content of social media posts, and to create a predictive model which would predict variables of interest from dependent variables. By using public data from Twitter or mental health forums large sample sizes were possible, but also resulted in less reliable data. It is estimated that the ability to detect depression by machine driven predictive models running on big social media data was above that of unaided primary care clinicians, but below that of self-report surveys. [5].

Clinical assessment and diagnostic assignment follows screening. There is considerable interest in developing more effective diagnostic assessment using big data analytics. Models were created using techniques such as data mining, machine learning, and natural language processing to group people into diagnostic categories based on data from a variety of sources. [12]. In bipolar research, machine learning algorithms looked for patterns in neuroimaging, genetic analysis, neuropsychological tests, and protein biomarkers. They were able to create predictive models, but their performance was not greater than current diagnostic systems. While this task could not be completely automated via big data analytics any time soon, it may inform clinical diagnosis in the short-term [8].

Predictive models using machine learning techniques are also being constructed from a variety of data sources to estimate patient outcomes, which could be helpful in selection of interventions at the onset of treatment [12]. Predictive risk profiles for patient's with bipolar disorder were created by taking data from Electronic Medical Records and identifying patient characteristics connected to negative outcomes, such as relapse and hospital admission. Studies also explored models which predict patient mood states, based on past monitoring data and how patients will respond to specific interventions. While these examples were fairly accurate (68% to 99%), they were based on relatively small sample sizes [8]. Predictive models show promise of being an effective big data application in mental health treatment, but require further advances in machine learning techniques and validation on larger samples before they can be widely administered [12].

2.2 Interventions

Once a person's mental health issues have been clinically identified, then interventions are assigned. Tradtional interventions are clinician driven, and are often limited in scope by clinician availability. Web-based interventions, which provide treatment activities via web-browser, have the potential to provide more flexible treatment options for patients. Initial attempts have seen some success, particularly if paired with a human coach. Few estimates of effectiveness exist, as these techniques have not been applied to large groups [10]. While big data approaches are not widely utilized, there is interest in using machine learning to predict content that a particular user would find helpful [11],which is a technique called a recommender system [17]. Also, as interactive interfaces are developed and used by large numbers of online users [10], big data analytics would be beneficial.

2.3 Treatment Monitoring

As a person receives treatment, tracking progress towards their goals is critical. Traditionally this is done by patient report via a tracking log or by clinician inquiry during a session, and is often hindered by a lack of patient engagement. One solution to this is active monitoring utilizing mobile devices. Patients can receive text messages or application notifications containing treatment goal reminders, symptom assessment questions, or encouraging messages to foster treatment participation [11]. Feedback from the patient can come in various forms from filling out a survey to voice response, and may be collected multiple times a day. The frequent collection of different types of data make active monitoring an application which could benefit from a big data approach. However, trouble with integrating data into the electronic medical record and a lack of widespread utilization have prevented such approaches from being extensively applied or reliably tested [12].

Another possibility is passive monitoring, which would access information from a mobile device, and connect those to patient behaviors, without any intentional action on the patient's part. This has been done using clinically informed algorithms or machine learning paired with self-report [11]. Devices used were not just smartphones, but including wearables and a sensor which is swallowed to detect medication adherence. Active monitoring has generated considerable research interest, but implementation at a big data level is challenged by lack of client engagement, clinician's ability to use, and difficulties integrating the large quantities and varieties of data [12].

3 DISCUSSION

3.1 Barriers

Overall, there is considerable interest in developing big data applications at every stage of the mental health process. However, this development has been slow and halting due to a number of issues inherent though not necessarily unique to human services.

For example, the issue of privacy is relevant with many big data applications, but in mental health the sensitive nature of an individual's mental health treatment data creates new difficulties. Typically privacy is preserved through de-identification of the data, but this is not always effective with large-scale data [12]. A specific privacy risk is big data analysis of social media, which captures large amounts of information, which can be used to infer mental health status [5]. When mental health privacy is breached, discrimination regarding employment, insurance, housing, etc. are possible [12]. On the other side of the privacy question, mental health professionals are mandated to report if someone is an imminent risk to themselves or others. Currently, there are no clear guidelines to follow, if this is discovered through public data [5].

Another challenge to capitalizing on big data is the variety of data sources, formats, and storage locations. The vast majority of mobile devices are not run on open source software, as they are sold as commercial products. This hinders collaboration and integration of the data with sources from other companies' products [13]. It is also unclear who owns the data in these situations, causing more disruption [12]. This is not just the case with private data. Large databases and research institutions often struggle to share data, and the decision to do so is often up to the individual researchers. This prevents the collaboration and coordination required to make good use of the available big data opportunities [6].

3.2 Future Directions

Considerable attention is being given to big data applications in mental health treatment, and some major initiatives seek to address some of the technical issues mentioned previously. The National Institute of Health's Office of Behavioral and Social Sciences Research has a strong focus on big data in its 2017 to 2021 strategic plan. It specifically called for the development of "data infrastructure that promotes data sharing, harmonization, and integration", and also to develop research methods which are designed for science which extensively uses big data [15]. There is a related call for treatment to inform research questions, and research questions to inform the structure and collection of big data, as opposed to primarily opportunistic research, which studies data that is most convenient [19]. The integration of private commercial data for big data analytics is also a goal of some researchers [13]. Concerning specific technologies, there is generally great optimism that the big data analytics techniques will continue to be refined, and that wider implementation will result in greater strides in treatment effectiveness.

Most of the research reviewed ended with a short description of ethical concerns in big data use for mental health treatment, and a call for someone to look into this in more detail. The problem is that there is a wide variety of perspectives about this topic. Some operate from the assumption that if data is publicly accessible, that resolves any privacy issues. Others point out cases where individual's privacy was seriously compromised by comparing data from multiple public databases [9]. This is a point where public policy has fallen behind technological innovation. An inter-disciplinary effort from legal, data science, and mental health experts may be required to strike the balance between science and citizen security [7].

4 CONCLUSION

At every stage, mental health treatment is a data intensive task. As electronic medical records, social media, and mobile devices continue to increase in data collection and storage capabilities, data relevant to mental health continues to grow larger, faster, and more varied. Many researchers and practitioners are eager to use big data analytics to tap into the potential insights of these data sets.

The first steps of development have already started, and show promise of making a significant positive impact in the field. Predictive analytics are being tested to screen for people with mental illness via social media, and machine learning techniques are being applied to improve the resolution of diagnosis and to inform treatment assignments through outcomes prediction. Though these results need replication with larger samples, they already demonstrate predictive power, which could soon equate with improved treatment in practice.

Applications utilizing mobile devices for active and passive monitoring of treatment participants are generating considerable attention, but are only early in development. As this approach is expanded to larger samples, big data analytics will be critical to managing the velocity and variety of data coming from smartphones and wearables. Integrating big data analytics in web-based mental health interventions, is even earlier in development. The potential to create interactive interfaces, utilizing artificial intelligence and recommender systems is present, but currently web-based treatments are being tested themselves for viability.

While progress to develop algorithms and programs to process mental health big data continues, it is hindered by the current limitations of data infrastructure and research culture. Though large data sources are available, they are not integrated with one another, and are often prevented from doing so due to preferences of individual researchers or from corporate interest. The National Institute of Health and many researchers are calling for an integrated and open data sharing framework to address this issue.

Also of concern is a variety of ethical questions involved in applying big data analytics to mental health. Ownership of data is not well defined, and often data is sold and studied without the knowledge of its subjects. During this process, an individual's privacy may be compromised, even with de-identified data. This can lead to discrimination and stigma for the individual whose mental health data has been unmasked. While this problem is readily recognized, no major policy or legislative change has adequately addressed it.

As big data analytics continues to mature, mental health treatment should seek to benefit from the unlocking of new knowledge and insights. However, this cannot be done without consideration of how to create an environment that simultaneously encourages practice innovation and patient protection. Treatment seeks to provide effective help to those with mental illness, and big data may help with that aim, but to do this at the expense of the patient rights undermines any help they hoped to gain.

ACKNOWLEDGMENTS

The researcher would like to thank Professor Gregor von Laszewski, along with Teaching Assistants Juliette Zerick, Saber Sheybani Moghadam, and Miao Jiang, and the anonymous reviewers who helped with the present work.

REFERENCES

[1] APA Practice Organization. 2017. Distinguishing Between Screening and Assessment for Mental and Behavioral Health Problems. Webpage. (2017). www.apapracticecentral.org/reimbursement/billing/assessment-screening.aspx

[2] Alex E Crosby, Beth Han, LaVonne A G Ortega, Sharyn E Parks, and Joseph Gfroerer. 2011. Suicidal thoughts and behaviors among adults aged ft18 years–United States, 2008-2009. *Morbidity And Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)* 60, 13 (2011), 1 – 22. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=22012169&site=eds-live&scope=site

[3] 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

[4] Jessica D. Goodman, James R. McKay, and Dominick DePhilippis. 2013. Progress monitoring in mental health and addiction treatment: A means of improving care. *Professional Psychology, Research and Practice* 44, 4 (2013), 231. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsggo&AN=edsgcl.354463723&site=eds-live&scope=site

[5] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43 – 49. http://proxyiub.uits.iu.edu/login?url=https://search-ebscohost-com.proxyiub.uits.iu.edu/login.aspx?direct=true&db=edselp&AN=S2352154617300384&site=eds-live&scope=site

[6] Diego Hidalgo-Mazzei, Andrea Murru, Mara Reinares, Eduard Vieta, and Francesc Colom. 2016. Big Data in mental health: a challenging fragmented future. *World Psychiatry: Official Journal Of The World Psychiatric Association (WPA)* 15, 2 (2016), 186 – 187. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=27265716&site=eds-live&scope=site

[7] Sharona Hoffman. 2015. CITIZEN SCIENCE: THE LAW AND ETHICS OF PUBLIC ACCESS TO MEDICAL BIG DATA. *Berkeley Technology Law Journal* 30, 3 (2015), 1741 – 1806. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=115393178&site=eds-live&scope=site

[8] Diego Librenza-Garcia, Bruno Jaskulski Kotzian, Jessica Yang, Benson Mwangi, Bo Cao, Luiza Nunes Pereira Lima, Mariane Bagatin Bermudez, Manuela Vianna Boeira, Flvio Kapczinski, and Ives Cavalcante Passos. 2017. The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews* 80 (2017), 538 – 554. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edselp&AN=S0149763417300337&site=eds-live&scope=site

[9] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211. https://doi.org/10.1177/2053951716650211 arXiv:https://doi.org/10.1177/2053951716650211

[10] Thomas D. Meyer, Rebecca Casarez, Satyajit S. Mohite, Nikki La Rosa, and M. Sriram Iyengar. 2018. Novel technology as platform for interventions for caregivers and individuals with severe mental health illnesses: A systematic review. *Journal of Affective Disorders* 226, Supplement C (2018), 169 – 177. https://doi.org/10.1016/j.jad.2017.09.012

[11] David C. Mohr, Michelle Nicole Burns, Stephen M. Schueller, Gregory Clarke, and Michael Klinkman. 2013. Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry* 35, 4 (2013), 332 – 338. https://doi.org/10.1016/j.genhosppsych.2013.03.008

[12] Scott Monteith, Tasha Glenn, John Geddes, Peter C. Whybrow, and Michael Bauer. 2016. Big data for bipolar disorder. *International Journal of Bipolar Disorders* 4, 1 (11 Apr 2016), 10. https://doi.org/10.1186/s40345-016-0051-7

[13] Andrea Murru, Eduard Vieta, and Francesc Colom. 2016. Big Data in mental health: a challenging fragmented future. *World Psychiatry* 15, 2 (2016), 186. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsoaf&AN=edsoaf.d1b40726336430fcecd614a54ac6d4b079039a17&site=eds-live&scope=site

[14] National Institute of Mental Health. 2017. (2017). https://www.nimh.nih.gov/health/statistics/index.shtml

[15] William T Riley. 2017. Behavioral and social sciences at the National Institutes of Health: Methods, measures, and data infrastructures as a scientific priority. *Health Psychology: Official Journal Of The Division Of Health Psychology, American Psychological Association* 36, 1 (2017), 5 – 7. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=28045300&site=eds-live&scope=site

[16] Social Security Administration. 2013. (2013). https://www.ssa.gov/policy/docs/statcomps/di_asr/2013/di_asr13.pdf

[17] Stanford Info Lab. 2017. Recommendation Systems. webpage. (2017). http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

[18] Z. Steel, C. Marnane, C. Iranpour, Tien Chey, J. W. Jackson, Patel Vikram, and D. Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980-2013. *International Journal of Epidemiology* 43, 2 (2014), 476 – 493. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=lhh&AN=20143278163&site=eds-live&scope=site

[19] Robert Stewart and Katrina Davis. 2016. 'Big data' in mental health research: current status and emerging possibilities. *Social Psychiatry and Psychiatric Epidemiology* 51, 8 (2016), 1055. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.460296829&site=eds-live&scope=site

[20] Substance Abuse and Mental Health Services Administration. 2015. (2015). https://www.samhsa.gov/data/sites/default/files/2015_National_Mental_Health_Services_Survey.pdf

[21] Substance Abuse and Mental Health Services Administration. 2017. Behavioral Health Treatments and Services. (2017). https://www.samhsa.gov/treatment

Big Data Applications and Analysis in Maternal Death During Childbirth in the United States

Elena Kirzhner
Indiana University Bloomington
3209 E 10th St
Bloomington, Indiana 47408
ekirzhne@iu.edu

ABSTRACT

Maternal mortality rate in the United States had increased by more than 25 percent from 2000 to 2014. Reducing maternal death during childbirth requires in-depth examination of isolated causes of death. With the major growth of big data and applications, it is possible to collect, analyze and compare specific maternal death causes and contributing factors to predict who's susceptible to fatality and what can be done to prevent it. It will help to develop focused clinical and public health prevention programs.

KEYWORDS

i523, hid320, Big Data Applications and Analytics, Data Science, Maternal Mortality

1 INTRODUCTION

Maternity death is rising for unclear reasons in United States. USA is the only developed nation where that rate is increasing and getting worse.

American women are more likely to die from childbirth than women in any other high developed country. Based on research and analysis by the Center for Disease Control and Prevention [1], maternal death greatly increased from 2000-2014 and more than half of such incidents could have been prevented with the current medical technology.

Most of the cases were result of medical error and unprepared hospitals. Doctor's ability to protect the health of mothers in childbirth is a basic measure of a society's development. Yet every year in the United States 700 to 900 women die from pregnancy or childbirth-related causes, and some 65,000 nearly die. By many measures, the worst record in the developed world [17] and [12].

We have ability to prevent it, by analyzing each cause and predict with monitoring the cases and usage of the Big Data and Analytics.

Statistical research for 2010 put America in the 50th place; the lowest of all developed nations for maternal death during childbirth[2]. Figure 1 shows Maternal Mortality ratio by developed countries per 100,000 live births [14].

From 1990 to 2014 pregnancy related death increased by 1.7% while worldwide that rate decreased by 1.3%. Thus, proper calculation shows that maternity mortality rate practically doubled in the last decade.

Figure 2 shows percent change in maternal deaths per 100,000 live births, from 1990-2013 [13].

Women giving birth in Asia have lower risk to die than those giving birth in United States [17].

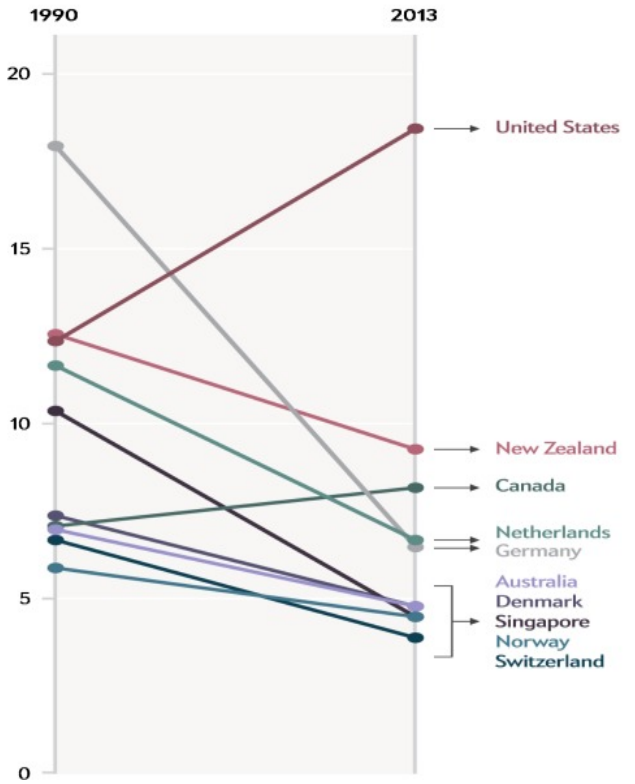

Figure 1: A comparison of maternal mortality ratio in the United States with those of some developed countries between 1990 and 2013 [14].

Currently, researches are inconclusive, as to why the rate is rising in USA. Multiple variables are being taken into account, such as race, age and economic status [5].

1.1 Definition

According to the National Center for Health Statistics, Pregnancy Mortality Surveillance System and the International Classification of Disease, to properly analyze data, causes of death during child birth were categorized and defined [3] as follows:

1. Pregnancy related death - death during the first 42 days after giving birth that is directly related to pregnancy and health care. Not related to any accidents outside of the pregnancy.

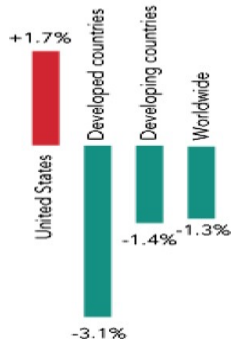

Figure 2: Percentage change in Maternal Mortality Rate between 1990 and 2013 in the United States, worldwide, developed and developing countries [13].

Figure 3: Changes in pregnancy related mortality ratio in United States from 1987-2011 [8].

2. Maternal fatality ratio - death caused by pregnancy for every 100,000 pregnancy occurrences.

1.2 Monitoring

The National Center for Health Statistics requires all states on annual basis to provide death certificates with causes of maternal death. This data is analyzed and compared against international statistics [11] and [4].

Additionally, Pregnancy Mortality Surveillance System was implemented in 1896, because of limited pregnancy death related records [10]. This system was created to record and analyze all pregnancy related deaths. Every year, this group sends a request to all 50 states to provide death certificate copies for those who died during childbirth and pregnancy. This data is stored and further analyzed by trained doctors, specialists and data scientists. That group coined a new term "pregnancy-related mortality" [3]. This information is being released in Center for Disease Control and Prevention Morbidity and Mortality Weekly reports and their website [16]. Deaths related to pregnancy from 1998-2010 were published in Obstetrics and Gynecology journal [18]. Furthermore, since launching the program, monitoring and analyzing the data, rate has dramatically increased from 7.2 deaths per 100,000 births in 1987 to 17.8 deaths per 100,000 births in 2011 [16]. Figure 3 shows changes in pregnancy related mortality ratio in United States from 1987-2011 [8].

2 BIG DATA USAGE AND HOW IT CAN HELP

The maternity deaths cases are well suited for a big data usage and solution. We have large amount of unstructured data. It could be used on advanced level for further examination. The data could be simplified and accessible to everyone including patients and doctors. However, there are not enough expertise to use it and limited financial resources. On the other hand, some experts say that it is good to have unstructured or raw data, because it was not modified and could add real value to healthcare analytics. It is fresh and ready to be consumed [6].

The causes of maternity death are not yet identified. We have only limited amount of data that was analyzed [5].

Moving forward, we need to understand and organize pregnancy related deaths and causes. Figure out structure and and identify risks by race ethnicity, economic status and age. Professional examination and generated analysis of structured and unstructured data could help with preventing causes of pregnancy related death.

2.1 Who Is Already Doing It

Over 200 healthcare applications were developed since 2010. Number of healthcare providers have already benefited from big data by concentrating on the fundamental structure of the big data. Few examples below:

Kaiser Permanente adapted new system called HealthConnect, it communicates new data between collected information about patients and treatments. The implemented system have helped to save more than one billion dollars from lowering patients visits to doctor's office [9].

Blue Shield of California adapted NantHealth and improved outcomes between patients and hospitals by communicating information about the visits, patient health history and hospitals. It helped to provide most effective and cheaper treatments for chronicle illness with preventive care and communications between doctors and patients [9].

The Lancet Journal done similar study on October 8, 2016 that called "Global, regional, and national levels of maternal mortality, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015" [13]. They used a standardized process to identify, extract and process all relevant data sources. Uniformed algorithms were applied to identify age category, year category, and location

specific patterns of failure and hidden records for vital registration, as well as patterns of deaths misrepresentation [15].

2.2 Internet Of Things

Big Data and the Internet Of Things is a growing system that allows convergence of physical equipment that transfers data and communicate with other devices and digital networks tearing down silo walls between operational technology and information technology. In conjunction with big data it allows for extraction of valuable information.

It could be used to monitor patient's health and their pregnancy risks such as diabetes level or blood pressure. It could also track prescribed medicine, it is especially useful for patients without health insurances [13].

2.3 Predictive Analytics

Many tools are being utilized for predictive analytics usage, such as data mining, statistics and historical facts. It is being used to analyze given information and generate predictions and outcomes for future and unforeseen events. Stored data could be useful, pregnant women's information could be shared between doctors and hospitals to be diagnosed in advance, improving number of healthy pregnancies. By being able to analyze relevant data, pregnancy risks could be predicted and provide women with safer and better pregnancy outcomes. The more analyzed data we have, the sooner it will reduce the mortality rates and we'll be able to diagnose each case. Special emergency kits with appropriate medicine could be supplied to each hospital and doctor's office for individual patient.

Huge amount of data is being generated daily and it comes from different sources in variety of shapes and sizes. Pregnancy related issues are being collected through social media, forums, blood tests, pharmaceutical companies, doctor visits, ultrasounds, hospitals, emails and so on. Our life became very digital. Currently, every doctor's visit is being recorded digitally, and electronically health records are being stored at health-care insurance departments and hospital facilities. These records are playing important part of research and scientific analysis.

2.4 Crunching Big Data

US government is focusing on research and transforming healthcare knowledge. Big data software becoming accessible and being developed for efficiency and made it easier to collect and analyze data from different sources. One of the best options for the data analysis is to input it into Hadoop system to make a more scaleable analysis with that. As of today, it is one of the most popular data management option. Additionally, it is one of the largest systems that is being used by many companies. Its ability to handle multiple amount of data from different sources, makes it productive and provides possibility to get more accurate causes and reasons of any health issues. Hadoop system is an open source software for distributed storage of large datasets on computer clusters and visualization. There are two main features; Hadoop Distributed File System, which responsible for files storage, and MapReduce, which generates and processes the data. The primary function of this programs is the capability to process huge amount of unstructured data and print out analyzed information. This system is all about handling the Big Data [7].

3 CONCLUSION

Pregnancy-related mortality findings should be studied and cross analyzed with the latest and advanced technology. It will provide a new view and value, resulting clarification and better health management.

Additionally, it will decrease same errors and doctors faults and prevent maternity death and its causes.

All these years, there was not enough information that was structured for deeper understanding and analysis. It can be improved. Big Data massively grows daily, useful information is everywhere around us; including emails, doctor's notes, lab tests, health insurances, ultrasounds, social media and pharmaceuticals .

Latest and fastest platforms such as Hadoop, have the ability to transform and improve the healthcare, store data and analyze huge mass of information from separate sources.

Doctors, medical staff and patients could use that information to improve and achieve better outcomes for pregnant mothers and prevent death. In addition, it will lower medical costs.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and Miao Jiang for their help, support and suggestions to write this paper.

REFERENCES

[1] SJ Bacak, CJ Berg, J Desmarais, E Hutchins, and E Locke. 2006. State maternal mortality review: Accomplishments of nine states. *Atlanta: Centers for Disease Control and Prevention* 0, 0 (2006), 1.

[2] Debra Bingham, Nan Strauss, and Francine Coeytaux. 2011. Maternal mortality in the United States: a human rights failure. (2011).

[3] William M Callaghan. 2012. Overview of maternal mortality in the United States. In *Seminars in perinatology*. Elsevier, Callaghan, William M, 0, 2–6.

[4] Andreea A Creanga, Cynthia J Berg, Jean Y Ko, Sherry L Farr, Van T Tong, F Carol Bruce, and William M Callaghan. 2014. Maternal mortality and morbidity in the United States: where are we now? *Journal of Women's Health* 23, 1 (2014), 3–9.

[5] Andreea A Creanga, Cynthia J Berg, Carla Syverson, Kristi Seed, F Carol Bruce, and William M Callaghan. 2012. Race, ethnicity, and nativity differentials in pregnancy-related mortality in the United States: 1993–2006. *Obstetrics & Gynecology* 120, 2, Part 1 (2012), 261–268.

[6] P Dineshkumar, R SenthilKumar, K Sujatha, RS Ponmagal, and VN Rajavarman. 2016. Big data analytics of IoT based Health care monitoring system. In *Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on (0)*. IEEE, 0, 0, 55–60.

[7] Jens Dittrich and Jorge-Arnulfo Quiané-Ruiz. 2012. Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2014–2015.

[8] Centers for Disease Control, Prevention, et al. 2014. Pregnancyrelated mortality surveillance. 2013. (2014).

[9] Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. 2016. The'big data'revolution in healthcare: Accelerating value and innovation. *0* 0, 12 (2016), 2014–2015.

[10] Isabelle L Horon and Diana Cheng. 2011. Effectiveness of pregnancy check boxes on death certificates in identifying pregnancy-associated mortality. *Public Health Reports* 126, 2 (2011), 195–200.

[11] Donna L Hoyert. 2007. Maternal mortality and related concepts. *Vital & health statistics. Series 3, Analytical and epidemiological studies/[US Dept. of Health and Human Services, Public Health Service, National Center for Health Statistics]* 0, 33 (2007), 1–13.

[12] Amnesty International. 2010. *Deadly Delivery: The Maternal Health Care Crisis In the USA*. Amnesty International Publications, 0.

[13] Nicholas J Kassebaum, Ryan M Barber, Zulfiqar A Bhutta, Lalit Dandona, Peter W Gething, Simon I Hay, Yohannes Kinfu, Heidi J Larson, Xiaofeng Liang, Stephen S Lim, et al. 2016. Global, regional, and national levels of maternal mortality, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* 388, 10053 (2016), 1775.

[14] Dina Fine Maron. 2015. Has maternal mortality really doubled in the US. *Scientific American* 0, 0 (2015), 0.
[15] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. 2013. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 0.
[16] Yasmin H Neggers. 2016. Trends in maternal mortality in the United States. *Reproductive Toxicology* 64 (2016), 72–76.
[17] World Health Organization, UNICEF, et al. 2012. Trends in maternal mortality: 1990 to 2010: WHO, UNICEF, UNFPA and The World Bank estimates. *0* 1 (2012), 10.
[18] Kenneth F Schulz, Iain Chalmers, David A Grimes, and Douglas G Altman. 1994. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *Jama* 272, 2 (1994), 125–128.

The Impact of Big Data on the Privacy of Mental Health Patients

J. Robert Langlois
Indiana University, School of Informatics and Computing
P.O. Box 1212
Bloomington, Indiana 47408, USA
langloir@umail.iu.edu

ABSTRACT

Today, one of the essential functions of technology is the collection, storage, processing, and transmission of data. The healthcare industry, including mental health services, are huge benefactors of these advances in technology. From birth, medical facilities start collecting information about all individuals; they do so even up to the point of death and all points in between. Over a lifetime, that is an abundance of information about an individual. The question that must be answered is, how is that data be protected to ensure patients' privacy rights? The more information collected on individuals, the more responsibility is assumed by those who collect data; methods for how the data is collected, used and shared must ensure the protection of patients' privacy rights. This challenge is one that needs to be navigated and addressed by medical professionals and facilities, policymakers, and the individuals whose data is collected. Specifically, in the mental health field, by resolving patients' privacy concerns, policymakers and researchers can transform the field by introducing more cost effective strategies, ensuring patients' sense of security, and establishing new and more appropriate norms to communicate sensitive health information.

KEYWORDS

i523, HID325, Big Data, Mental Health, Privacy

1 INTRODUCTION

We live in an era where data is constantly being produced; data exists everywhere in large quantities. The advances in technology have opened the door for businesses to collect inconceivable amounts of information on individuals via emails, smart-phones, sensors, and other technology devices. The 21st century has witnessed a data explosion; many fields have experienced a data deluge that can contribute to boast the economy via data analysis, make new discoveries based on existing data, respond to health problems in a quicker manner, and so forth. While it is worth celebrating the rapid innovations in technology and the presence of huge amounts of data, it is also crucial to consider the number of barriers and risks that come with the increased availability of data; often refers to as big data.

One of the barriers that big data faces is privacy. In the healthcare industry, for example, there are protocols to accessing data that can cause financial burdens and can be time-consuming. The cost of collecting, disseminating, and organizing patient information, along with the time it takes to handle the information are some of the challenges. There are also very serious concerns regarding who can have access to what kind of patient information. Policymakers have a very important role in establishing more up-to-date policies and parameters that address the massive amounts of information available and the appropriate ways to collect, share, and house the data. "When considering the risks that big data poses to individual privacy, policymakers should be mindful of its sizable benefits" [7]. While it is important to address the numerous advantages of big data, it remains relevant to figure out ways to prevent data leakage, and to protect the privacy of individuals. This paper showcases the advantages of big data and the ways to overcome the individual privacy concerns.

2 THE ADVANTAGES OF BIG DATA

Big data analysis presents numerous advantages. For instance, it helps businesses to increase their productivity. This has done through a process of analyzing raw data that produces information that identifies trends and patterns that will help businesses make cost effective decisions. It is also helpful in aiding government agencies to improve public sector administration, and assists global organizations in analyzing information that has wide-reaching impact on the world. The information produced by big data can help medical professionals to detect diseases in earlier stages. Some other advantages of big data analysis is present in many different areas, such as: smart grids, which monitor and control electricity use; traffic management systems, which provide information about transportation infrastructure likes roads and highways, mass transit, construction, and traffic congestion; retail by studying customer purchasing behavior to improve store layout and marketing; payment processing by helping to detect fraudulent activity, etc [7].

Certain research studies have supported the idea that big data allows for real time tracking of diseases and the development, prediction of outbreaks, and facilitates the development of personalized healthcare. Big data can also be used to maximize profits in many disciplines, including healthcare if harnessed properly [8]. "by harnessing big data, businesses gain many advantages, including increased operational efficiency, informed strategic direction, improved customer service, new products, and new customers and markets" [2]. While data exists in huge quantities in many fields, including the health care field, individual privacy concerns remain a big problem that policymakers have to tackle to meet current trends in data collection. Improved methods of protecting very personal, private and sensitive health information is needed in order to allow for safe, necessary and adequate access to protected health information within the health care industry. Without proper policies related to data use, access, and protection, this big data potential can not be realized [4]. What are the barriers to big data in heathcare?

3 THE BARRIERS TO BIG DATA IN HEALTH-CARE

One of the barriers faced by big data analysts in healthcare, including mental health services, is privacy. Regardless of the efforts policymakers try to establish, the different strategies in place to protect individual health information can pose serious challenges that scientists have to wrestle with when it comes to big data analytics. One of the most notable efforts that policymakers have introduced to secure health information, is the creation of the Health Insurance Portability and Accountability Act (HIPAA) in 1996. HIPAA has established norms for data privacy and has mandated security provisions for safeguarding medical and mental health information. Every provider in the healthcare industry must comply with HIPAA privacy laws if they want their practices to remain up and running. The HIPAA laws prohibit providers from sharing patients' information without their consent. The challenge for big data analysts is that a lot of times, patients refuse to share their personal information for research purposes due to fears that the health issue will be the cause of being ostracized, discriminated against, marginalized, etc. "The unintended release of a person's health information into the public realm has huge potential to undermine personal dignity and cause embarrassment and financial harm" [8]. While the healthcare field is faced with a huge increase in health information, individual privacy concern remains a huge conundrum for big data analysis. What can policymakers do to overcome individual privacy concerns, but still allow for the sharing of information that would be for the better good of society at large?

4 WAYS TO OVERCOME PRIVACY CONCERN

4.0.1 Data Anonymization. One way policymakers can protect individual privacy is by making the data anonymous. Researchers have identified three types of data: personal and proprietary data that is controlled by individuals; government-controlled data, which government agencies can restrict access to; and, open data commons, which means that the data is centrally located and available to all. Big data analysts and researchers have advocated for linking data together that can help to improve health care planning at both the patient and population levels. They also argued for an increase in the amount of information that is available in open data commons [4]. Although the anonymization of data appears to be a great technique that policymakers could espouse to address privacy concerns, other studies have indicated that some data can be traced back to their respective individual; thus, destroying the argument for anonymity [8]. "Every copy of data increases the risk of unintended disclosure. To reduce this risk, data should be anonymized before transfer; upon receipt, the recipient will have no choice but anonymize it at rest...And re-identification is by design, in order to ensure accountability, reconciliation and audit" [1] If proper norms are established for data analysis, this can potentially contribute to improvements in the health care industry.

Still, there are others that have advocated for data de-identification and data minimization. The term de-identification is the process by which the data is made anonymous. The proponents of this process explain that this protective measure is valid under security and accountability principles, but admonish that policymakers should think about other ways to protect patients' privacy. The term data minimization, describes the extent to which organizations can limit the collection of personal data. It is worth noting that data minimization is contrary to big data analysis because data minimization encourages deleting data that is no longer in use in order to protect privacy; whereas, big data analysts would prefer to archive the data for ulterior usage. While this technique can help protect privacy, it is antithetical to big data analysis because it contributes to reducing the amount of data collection that could be used in data analysis to make new discoveries, respond to crises, and maximize profits [7].

Privacy principles should be introduced during the process of data architecture; privacy should be incorporated into the design and operational procedures [1]. In so doing, personal health care data will be protected against malicious hackers who try to access individuals' personal health information for the purposes of stealing individuals' identity. Another type of data that has been introduced to the healthcare industry is concept quantified self data. It can be understood as the data produced by individuals that engage in self-tracking of personal health information, such as heart rate, weight, energy levels, sleep quality, cognitive performance, etc. These individuals use devices like smart-phones, watches, and wearable technology sensors in the collection of their personal data and biometrics. It has been shown that 60 percent of U.S. adults are tracking their weight, diet or exercise routines, while 33 percent are monitoring their blood sugar, blood pressure, sleep patterns, etc. This indicates that there is a vast amount of health information that has been produced by individuals. What is done with all of this data? This massive supply demonstrates the need to develop policies and protocols that involve individual patient consent to share their collected data; this data can be critical to the advancement of healthcare with the support of data analysis. Before that can be done, however, we must first establish the proper norm to use this type of data so that the privacy of individuals can be protected; this ought to be the primary action to take. [6].

In the healthcare industry, Patients often do not want their health information to fall in the hand of other entities without their consent; however, with proper informed consent, patients seemed to become willing to share their personal health information.As agencies work with patients to disclose the purposes of collecting certain, sometimes sensitive, health information, they can empower patients to make informed decisions about their personal health information, thus engaging patients in the process. This can then serve to increase and improve the set of personal health information utilized for clinical research purposes, and subsequently improve people's lives [5]. "Privacy concerns exist wherever personally identifiable information or other sensitive information is collected and stored in any form" [3]. Thus, to protect privacy, other techniques, like encryption, authentication, and data masking may be utilized to ensure that the information is available only to authorized users.

5 CONCLUSIONS

We have seen that healthcare data exists in large quantities; however, privacy concerns are one of the biggest barriers and challenges that scientists face when it comes to utilization of healthcare data. Certain researchers have proposed data anonymization as a solution to privacy concerns, while others have proposed a minimization of

the amount of data collected on individual patients, as well as authenticate the data so that it can only be accessed by intended users. Suggestion was also made to involve patients in the collection of health data, so that they can be more willing to share their information that can play a vital role in improving healthcare and mental health research, reduce health care cost, maximize profits, etc. It is almost certain that scientists will always have to wrestle with privacy concern whenever they are dealing with personal health information; thus the importance for policymakers to continue to encourage dialogue among healthcare providers and patients, and develop policies and regulations on how to utilize healthcare data without compromising patients' privacy rights.

REFERENCES

[1] Ann Cavoukian and Jeff Jonas. 2012. *Privacy by design in the age of big data.* Information and Privacy Commissioner of Ontario, Canada.

[2] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal* 2014 (2014), 18.

[3] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. 2016. Digital Health Data: A Comprehensive Review of Privacy and Security Risks and Some Recommendations. *Computer Science Journal of Moldova* 24, 2 (2016), 71.

[4] Joachim Roski, George W Bo-Linn, and Timothy A Andrews. 2014. Creating value in health care through big data: opportunities and policy implications. *Health affairs* 33, 7 (2014), 1115–1122.

[5] Robert H Shelton. 2011. Electronic consent channels: preserving patient privacy without handcuffing researchers. *Science translational medicine* 3, 69 (2011), 69cm4–69cm4.

[6] Melanie Swan. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data* 1, 2 (2013), 85–99.

[7] 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.

[8] J Van Den Bos, K Rustagi, T Gray, M Halford, E Zeimkiewicz, and J Shreve. 2011. Health affairs: At the intersection of health, health care and policy. *Health Affairs* 30 (2011), 596–603.

Big Data in Clinical Trials

Mohan Mahendrakar
Indiana University
P.O. Box 1212
Bloomington, Indiana 43017-6221
mmahendr@iu.edu

ABSTRACT

Will understand about Clinical Trials and how Big data is impacting Clinical Trials. Clinical Trials is experiencing a data-driven transformation. Clinical trials getting ready with new and ever extra efficient molecular and computer technologies, we are entering new era where big data technologies are helping us forefront the contest against various deceases. This technology driven information bang, often denoted to as "big data".

KEYWORDS

i523, HID326, Big data, Clinical, Trials, Health care, Data integration, Analytics

1 INTRODUCTION

A prime focus of clinical trials is gaining knowledge from studying a group of patients which can then be functional to a much wider group of patients to recover from deceases. In general practice, providing care to patients is delivered within a rich background of intrinsic and endemic confusing issues and prejudices related with practices and patients[2].

The data received from around the world from various patients, decease form Big data, Big data nothing but collection of large data sets. These data sets may grow even beyond petabytes in size. In clinical trials big data usage just started and but big data use cases are promising and widely used in near future. Few hail the great use cases and some are neutral about big data but big data is game changing technology in clinical trials[6].

According to [4] digital world grows at aspect of 300, to 40,000 exabytes from 130 exabytes. until year 2020 digital world keep doubled every 2 years.

According to [7] it is predicted that the market for Big Data technology and services will reach $20 billion in 2017, up from $3.2 billion in 2010. This is an annual growth rate of 40 percent, which is about seven times the rate of the overall information and communications technology market. According to CB insights, health care investments in Big Data totaled $274.5 million in 2012, and it went to $371.5 million in 2013.

2 BIG DATA & CLINICAL RESEARCH

Determining clinical trials hidden data patterns and relations within the mixed data, discovering new pharma companies and drug goals. Letting the new development of predictive disease progression models. Analyzing Real World Data (RWD) as a balancing instrument to clinical trials, for the rapid development of new personalized medicines. The expansion of progressive statistical methods for learning fundamental relations from large scale observational data is a very important factor in the analysis[5].

2.1 Data Integration

Having access to right, appropriate and trustable and associated is a biggest challenges facing medical clinical trials. The ability to accomplish and integrate data collected at all phases of the clinical trials right from detection to real world usage after regulatory approval, this is a essential goal to let organizations to originate more profit from big data technologies. Value addition analytics are designed on data as the foundation. The trusted sources of all pieces are formed based on effective end-to-end data integration and relates dissimilar data irrespective of source that can be internal or external, publicly available or patented. Another benefit of data integration is one can perform wide-ranging searches for subsections of information based on the relations established rather than only available data. "Smart" algorithms which connects to clinical trials information and laboratory could generate automatic reports that helps to find right applications or compounds and even generate flags which helps in understanding safety or effectiveness [2]. Applying data integration end-to-end needs a lot of competences, including but not limited certified sources of documents and data, the capability to create cross relations among the elements, robust quality assurance, workflow management, and accesses based on roles to safeguard that definite data elements are available only to authorized to access and see it. Medical organizations usually evade overhauling their entire data-integration systems at one point of time due to the logistical challenges and costs associated, even though there are few international pharmaceutical enterprises employing a "big bang" method to redesigning its clinical IT systems[2].

Data is being generated by different sources and comes in a variety of formats including unstructured data. All of this data needs to be integrated or ingested into big Data Repositories or Data Warehouses. This involves at least three steps, namely, Extract, Transform and Load (ETL). With the ETL processes that have to be tailored for medical data have to identify and overcome structural, syntactic, and semantic heterogeneity across the different data sources. The syntactic heterogeneity appears in forms of different data access interfaces, which were mentioned above, and need to be wrapped and mediated. Structural heterogeneity refers to different data models and different data schema models that require integration on schema level. Finally, the process of integration can result in duplication of data that requires consolidation[5].

The process of data integration can be further enhanced with information extraction, machine learning, and semantic web technologies that enable context based information interpretation. Information extraction will be a mean to obtain data from additional sources for enrichment, which improves the accuracy of data integration routines, such as duplication and data alignment. Applying

an active learning approach ensures that the deployment of automatic data integration routines will meet a required level of data quality. Finally, the semantic web technology can be used to generate graph based knowledge bases and oncologies to represent important concepts and mappings in the data. The use of standardized oncologies will facilitate collaboration, sharing, modelling, and reuse across applications[5].

2.2 Exascale computing

After data integration is completed, the big question is how to process such huge volume of the data? There will be use cases, e.g. precision medicine, where the promises brought by big data will only be fulfilled through dramatic improvements in computational performance and capacity, along with advances in software, tools, and algorithms. Exascale computers-machines that perform one billion calculations per second and are over 100 times more powerful than today's fastest systems will be needed to analyses vast stores of clinical and genomic data and develop predictive treatments based on advanced 3D multi-scale simulations with uncertainty quantification. Precision medicine will also require scaling these systems down, so clinicians can incorporate research breakthroughs into everyday practice[5].

2.3 Data-driven metamorphosis

Data collected in clinical trials experiencing a data driven metamorphosis. Information technologies equipped with new and even more efficient molecular, we are in the era where information is supporting us driving force to fight against to deceases like cancer. This expertise driven data blast, generally mentioned as "big data", is not only helping discoveries in biomedical, then it is also rapidly applying the practice of oncology into an information science. This development is very critical, as outcomes to-date have opened the enormous complication and genetic heterogeneity of trials patients and patients tumors, a soberingnotice of the challenge undergoing each patient and their oncologist. The answer to this issue is addressed only through developing data analytics in clinico-molecular, that will help deeper analyses of mechanics which is controlling the biological and clinical response to available therapeutic options. Beyond the available guidelines for better-quality patient care, such progressions in predictive and evidence-based analytic stand to deeply impact the existing processes in discovering the drugs for cancer drug and also corresponding clinical trials [3].

2.4 Big data analytics

Medical research has always been a data-driven science, with randomized clinical trials being a gold standard in many cases. However, due to recent advances in omics-technologies, medical imaging, comprehensive electronic health records, and smart devices, medical research as well as clinical practice are quickly changing into big data-driven fields. As such, the healthcare domain as a whole - doctors, patients, management, insurance, and politics - can significantly profit from current advances in Big Data technologies, and from analytics[5].

2.5 Machine Learning

Many healthcare applications would significantly benefit from the processing and analysis of multimodal data - such as images, signals, video, 3D models, genomic sequences, reports, etc. Advanced machine learning systems can be used to learn and relate information from multiple sources and identify hidden correlations not visible when considering only one source of data. For instance, combining features from images (e.g. CT scans, radiographs) and text (e.g. clinical reports) can significantly improve the performance of solutions[5].

3 CHALLENGES

Large biomedical organizations typically save their discoveries confidential due to the costs associated in developing the drug throughout its life cycle almost 12 years it may take for a medication to be ready on prescription pad from discovery and also very costly deal about $4 billion to spent for the whole process, because of costly investments, it is not feasible option to share the secrets of upcoming blockbuster drugs, on top of it only ten percent of drugs finish its life cycle and come to market[1].

Although there is already a huge amount of healthcare data around the world and while it is growing at an exponential rate, nearly all the data is stored in individually. Data collected by a clinic or by a hospital is mostly kept within the boundaries of the healthcare provider. Moreover, data stored within a hospital is hardly ever integrated across multiple IT systems. For example, if we consider all the available data at a hospital from a single patient's perspective, information about the patient will exist in the EMR system, laboratory, imaging system and prescription databases. Information describing which doctors and nurses attended to the specific patient will also exist. However, in most of cases, every data source mentioned here is stored in separate silos. Thus, deriving insights and therefore value from the aggregation of these data sets is not possible at this stage. It is also important to realize that in today's world a patient's medical data does not only reside within the boundaries of a healthcare provider. The medical insurance and pharmaceuticals industries also hold information about specific claims and the characteristics of prescribed drugs respectively. Increasingly, patient-generated data from IoT devices such as fitness trackers, blood pressure monitors and weighing scales are also providing critical information about the day-to-day lifestyle characteristics of an individual. Insights derived from such data generated by the linking among EMR data, vital data, laboratory data, medication information, symptoms (to mention some of these) and their aggregation, even more with doctor notes, patient discharge letters, patient diaries, medical publications, namely linking structured with unstructured data, can be crucial to design coaching programs that would help improve people's lifestyles and eventually reduce incidences of chronic disease, medication and hospitalization[5].

4 CONCLUSION

The latest trends in big data creativities in health care is bringing confident Influence on clinical trials. Increased relations between collected data elements and nomenclature should help in streamline of trial designing and sharing of data. The process of standardization and quality improvement work go side by side with a growing

big data infrastructure applying guarantee benefits to information curation for trials.

ACKNOWLEDGMENTS

The authors would like to thank to Professor and TAs for guiding in making the better paper.

REFERENCES

[1] Jennifer Bresnick. 2014. *Big pharma opens up big data for clinical trials, analytics.* White Paper. Intelligent Media Network. https://healthitanalytics.com/news/big-pharma-opens-up-big-data-for-clinical-trials-analytics

[2] Jamie Cattell, Sastry Chilikuri, and Michael Levy. 2013. *How big data can revolutionize pharmaceutical R&D.* White Paper. McKinsey Center for Government. https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/public%20sector/regulatory%20excellence/how_big_data_can_revolutionize_pharmaceutical_research.ashx

[3] Taglang G and Jackson DB. 2016. *Use of "big data" in drug discovery and clinical trials.* Article. Molecular Health GmbH, 69115 Heidelberg, Germany. https://doi.org/10.1016/j.ygyno.2016.02.022

[4] John Gantz and David Reinsel. 2012. *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.* White Paper. EMC Corporation, 5 Speen Street Framingham, MA 01701 USA. https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[5] Dr. Adrienne Heinrich, Aizea Lojo, Dr. Alejandro Rodrguez Gonzlez, Dr. Andrejs Vasiljevs, Chiara Garattini, Cristobal Costa-Soria, Dirk Hamelinck, Elvira Narro Artigot, Prof. Ernestina Menasalvas, PD Dr. habil. Feiyu Xu, Dr. Felix Sasaki, Prof. Frank Mller Aarestrup, Gisele Roesems fi?! Kerremans, Jack Thoms, Marga Martin Sanchez, Marija Despenic, Mario Romao, Matteo Melideo, Prof. Dr. Miguel A. Mayer, Prof. Dr. Milan Petkovic, Dr. Nenad Stojanovic, Nozha Boujemaa, Patricia Casla Mag, Paul Czech, Prof. Roel Wuyts, Sergio Consoli, Dr. rer. Nat. Stefan Rping, Stuart Campbell, Dr. Supriyo Chatterjea, Prof. Dr. Ir. Wessel Kraaij, Wilfried Verachtert, Dr. Wouter Spek, and Ziawasch Abedjan. 2016. *Big Data Technologies in Healthcare.* techreport. Big data value association. http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20Healthcare.pdf

[6] F. Hoffmann-La Roche Ltd. 2013. *Understanding Clinical Trials.* techreport. GPS Public Affairs, 4070, Basel, Switzerland. https://www.roche.com/dam/jcr:1d4d1b52-7e01-43ac-862f-17bb59912485/en/understanding_clinical_trials.pdf

[7] Dr. Sarika Vanarse. 2014. *BIG DATA BREATHES LIFE INTO NEXT-GEN PHARMA R&D.* techreport. Wipro, DODDAKANNELLI, SARJAPUR ROAD, BANGALORE - 560 035, INDIA. http://www.wipro.com/documents/big-data-breathes-life-into-next-gen-pharma-RD.pdf

Using Big Data to Minimize Fraud, Waste, and Abuse (FWA) in United States Healthcare

Paul Marks
Indiana University
Online Student
Shepherdsville, Kentucky 40165
pcmarks@iu.edu

ABSTRACT

The cost of healthcare includes the loss of billions of dollars due to Fraud, Waste, and Abuse (FWA). Many of the schemes to commit FWA are very intricate and require the analysis of many data sources simultaneously. The question answered here is "How can we use big data analysis to help minimize these costs and thus optimize the money spent on healthcare?"

KEYWORDS

i523, hid327, fraud, waste, abuse, healthcare, health insurance

1 INTRODUCTION

FWA is an issue that affects everyone in the U.S. since healthcare services are leveraged by everyone at some point and the costs for those services include the money lost to FWA. The three components of FWA are varying degrees of culpability. The Centers for Medicare and Medicaid Services (CMS) in part defines fraud as "knowingly and willfully executing, or attempting to execute, a scheme or artifice to defraud any health care benefit program", Waste as "overusing services, or other practices that, directly or indirectly, result in unnecessary costs", and Abuse as "payment for items or services when there is not legal entitlement to that payment and the provider has not knowingly and/or intentionally misrepresented facts".[9] While the percentage of cost attributable to FWA can vary from insurer to insurer, Medicare estimates that 11 percent of its payments for Original Medicare are improper primarily due to FWA.[8] In combination these cost the United States healthcare system 80 billion dollars[6] annually.

Advances in big data technology can help reduce these losses. Big data offers the ability to look at data in real time to determine if a claim is legitimate or not. Historically, due to the amount of data involved, this type of analysis would have to happen after the claims have been paid with specific models targeting specific schemes to identify FWA. Big data can help lower the cost of health-care in the United States by identifying FWA claims and stopping payments before they occur.

2 HEALTHCARE FRAUD, WASTE, AND ABUSE ENVIRONMENT

It is easy to understand the problem FWA poses. Healthcare funds are of limited quantity. Insurance helps to spread the cost among groups of people, but does not provide limitless funds. As costs increase, so do premiums or direct payments for healthcare. In order for as many people as possible to be able to have access to healthcare costs have to be managed. There are many ideas for helping to provide affordable healthcare, but there is much discussion and disagreement on exactly how to do that. Reducing costs by eliminating as much FWA as possible is one solution that everyone, except for those participating in and profiting from FWA schemes, can agree on.

Data to fight FWA is not just the information gathered by a doctor or other provider while working with a patient. In order to fully utilize advances in technology, multiple sources of information must be brought together. Sources include claims (current and historic), clinical, provider, geospatial, and other sources of information. This allows for data analytics to take a deeper look into not only a single participant, but others who may be related to that participant. Providers who are involved in improper billing tend to be associated with other providers who have a higher frequency of being involved with improper billing as well. For big data analysis this involves looking at corporate ownership, who providers use as a billing agent, and who they tend to refer patients to for other service in order to uncover a larger pattern of FWA collusion.[13]

The problem for big data to solve is the size of all this data and how to process it fast enough. Using CMS as an example, being a government entity much of their data is available publicly, it is easy to get an idea of the amount of data. Medicare processed 1.2 billion claims in 2014, covering 53.8 million beneficiaries, with 6,142 hospitals, and 1,173,802 non-institutional providers[7]. In addition payments must be made within a specific timeframe depending on the insurer and their agreement with providers. This time includes all the normal steps to verify and process a claim so the time available to examine the data for FWA is very limited.

It must be noted that when working with this type of data, Protected Health Information (PHI) and Personally Identifiable Information (PII), that there are many regulations about the ability to access and secure it which must be followed. While this makes it more difficult to get access to the data it can be overcome by working cooperatively with the various data owners.

2.1 Big Data Techniques for FWA

So how can big data be used to approach this issue? Leveraging big data tools such as Hadoop, analysts could divide the different sources of information into data lakes, looking at each source separately, and then combining the results. Table 1 on page 3 shows sources of information and what level of FWA they are generally related to. The highest level combines sets of data from all data views. It looks for patterns across criminal networks which may involve many providers and beneficiaries. By looking at things more globally across potentially billions of records, big data provides the ability to perform complex network analysis which can uncover

intricate conspiracies perpetrated by coordinated efforts of many providers and facilities.[14]

While there are simple cases of fraud which follow a typical known pattern, this is only a portion of the problem. Fraud schemes change and can involve many different entities which may not seem to be related on the surface. The more data which can be combined and analyzed, the more fraud that can be found. The need for this type of analysis is because much of the FWA committed in healthcare is done so by providers working in conjunction with each other and providers working in conjunction with their patients.[5] Big data analytics can find hidden relationships and patterns in information which show FWA clusters. These can include, but are not limited to:

- Relationships between patients and the people who are committing fraud,
- Connections among those committing fraud, employees, businesses, and even their relatives,
- Suspect interactions between providers, and
- Overall inappropriate relationships among various active fraud participants, partners, and patients.[5]

In order to keep up with organized fraud activities, there must be a dedicated practice of data analytics which is ever evolving.

Traditionally programs have been written to look for specific sets of circumstances. Leveraging existing knowledge about the data and using it to look for specific patterns is known as supervised in big data terms. Supervised fraud detection is represented in several methods including "Bayesian Networks, Neural Networks (NNs), Decision Trees, and Fuzzy Logic."[3] Neural Networks and Decision Trees have a higher tolerance for handling large amounts of noisy data and are therefore more popular than the other methods. There are also unsupervised methods in which data is fed into the system without preexisting notions of what to look for[3]. Unsupervised methods sort through data and find relationships and groupings of related information, find clusters of what could be considered normal, and determine where the outliers are.

Because unsupervised methods only identify outliers, applying unsupervised methods to healthcare data requires that outliers then have to be verified as FWA or acceptable patterns. According to Anthem's SVP of Healthcare Analytics Patrick McIntyre, they take this into account. Anthem is able to run algorithms against their claims as they are being processed. This allows machine learning to discover claims which may be fraudulent or wasteful in nature on a daily basis. Once identified "questionable claims are immediately identified, flagged and sent to the clinical coding experts for review."[4] This greatly increases the ability to fight FWA by having the machine pinpoint where to look in all the data available to the reviewer. Suddenly the task of finding fraud is not as daunting. By leveraging both of these techniques FWA can be discovered at an accelerated pace. The number of models the system knows will grow over time as more data is fed into it and more patterns are discovered and verified.

2.2 Current Solutions

Many companies currently offer solutions for detecting FWA in healthcare payment systems. They include the ability to identify FWA claims during the payment cycle so that payment is not made to suspect claims. Truven Health[1], Healthcare Fraud Shield[12], and SAS[10], just to name a few, all have systems they offer based on big data. The specifics of the systems they offer are proprietary in nature so many of the descriptions are generic. Truven Health claims their solution mixes technology and healthcare intelligence. They have a model which groups services together for form a picture of the full view of the illness including inpatient, outpatient, and pharmaceuticals. This data in analyzed by knowledge rules based on clinical classifications and medical literature. This helps to identify wasteful or unnecessary service patterns in clinical and billing abuse which are hard to detect. Using this approach an analyst can look at the costs associated with the patient during their illness, the services provided, and combine this with others to form a profile of the provider's practice.[1]

SAS materials include the ability to find hundreds of millions of dollars in savings before claims are paid by taking an enterprise approach to FWA.[10] Their solution incorporates the rules and models into the claims process so companies can process the rules against all of their claims instead of sample sets. This helps to uncover more schemes and "spot linked entities and crime rings, which can help stem larger losses."[11]. Recently the Centers for Medicare and Medicaid Services awarded Northrop Grumman Corp. a contract worth $91 million to develop a second generation advanced analytics system to fight FWA in Medicare and Medicaid by identifying high-risk claims.[2]

2.3 Future uses of Big Data Analytics

Currently there is still a certain amount of honor built into healthcare. The inherent structure of the healthcare reimbursement system allows for both billing errors and fraudulent actors to go undetected, taking money away from the system set up to pay legitimate claims.[13] If a claim is submitted by a valid entity, using the correct process, and everything is in order then it is most likely paid. For many claims this is done without any specific proof of the services being provided. With more and more healthcare information being digitized this may not be the case in the future. X-rays, lab tests, clinical notes, etc. are all being stored digitally. Computers are now able to interpret images and unstructured text very accurately. By linking this data to claims data the clinical information could be required as part of claims payment. An x-ray of broken bone, notes which support a diagnosis, Magnetic Resonance Imaging files, could all be interpreted automatically. Not only would the data be used to compare to the claims information, but to other images/notes on file to ensure that the same files were not being submitted with multiple claims. The system could know what one individual medical history looks like compared to another similar to how facial recognition is able to match like images. Requiring and being able to validate more information before services are paid for would help the reduce the ability of perpetrators of FWA to be able to get reimbursed for services they should not. This level of verification would not be possible without the ability to process massive amounts of data quickly.

Historically the payers of most healthcare claims, insurers, have not had the ability to examine actual evidence that a service has taken place on a broad scale. (It is done manually on a specific case

Table 1: Types of Fraud and their related Sources[14]

		Phantom Billing	Duplicate Billing	Upcoding	Unbundling	Excessive or Unnecessary Services	Kickbacks
Level 1	Single Claim, or Transaction				*	*	
Level 2	Patient / Provider		*		*	*	
Level 3	a. Patient	*	***	*	***	*	
	b. Provider	**		***	*	***	
Level 4	a. Insurer Policy / Provider	**		*	**	**	*
	b. Patient / Provider Group	*	*	*	*	*	
Level 5	Insurer Policy / Provider Group	**		**	**	**	*
Level 6	a. Defined Patient Group	**		*	*	**	**
	b. Provider Group	**		***	**	***	*
Level 7	Multiparty, Criminal Conspiracies	**		**	*	**	***

Usefulness: * Low ** Medium *** High

or audit basis.) Through the use of advances in big data and combining current and new data stores such as electronic health records into the payment process, a difference can be made in the amount of money lost to FWA in healthcare. Data from providers must be run against entity resolution solutions. Complex claims analysis, including rules-based clinical reviews, must be part of the normal pre-payment workflow leveraging predictive analytics to stop payments on billions of dollars worth of FWA claims before they are made. The FWA problem goes beyond healthcare. It is "a national economic imperative that must be addressed immediately."[5] The technology exists today that can help protect the integrity of the healthcare system and the quality of care for Americans. [5]

3 CONCLUSIONS

While there may be disagreement on many aspects of healthcare in America, everyone should agree that eliminating Fraud, Waste, and Abuse within the system is the right thing to do. FWA costs billions of dollars annually. Just a 1 percent reduction in the estimated 80 billion dollars annually would result in 800 million dollars in savings. With this amount of money at stake significant investments should continue to be made in leveraging advanced big data technologies into solving this problem. Due to the continued rise in the amount of data collected traditional programming cannot keep up with the pace. Advanced techniques must be leveraged which can learn in an unsupervised manner. The future of the best methods for fighting FWA in healthcare will be a combination of this analysis and teams specializing in the rules and regulations of healthcare in the United States. The unsupervised methods will work through massive amounts of structured and unstructured data breaking it down into cases and schemes which are most likely FWA. These will be reviewed, confirmed or denied as accurate, and fed back into overall FWA platform. As this cycle continues over and over the ability to fight FWA in United States Healthcare will get better. While Big Data may never eliminate FWA in Healthcare it can help to minimize it and save the country billions of dollars a year.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper. It has helped to expand my knowledge in how modern data analytics can help to save FWA which has plagued the healthcare system.

REFERENCES

[1] Truven Health Analytics. 2017. Program Integrity. Online. (2017). http://truvenhealth.com/your-healthcare-focus/government/program-integrity

[2] Virgil Dickson. 2016. Northrop Grumman wins $ 91 million CMS contract for fraud detection. Online. (04 2016). http://www.modernhealthcare.com/article/20160407/NEWS/160409912

[3] Namrata Ghuse, Pranali Pawar, and Amol Potgantwar. 2017. An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques. *International Journal of Scientific Research in Network Security and Communication* 5, 3 (06 2017), 27–33.

[4] Erin Hitchcock. 2017. The Role of Big Data in Preventing Healthcare Fraud, Waste and Abuse. Online. (09 2017). https://www.datameer.com/company/datameer-blog/role-big-data-preventing-healthcare-fraud-waste-abuse/

[5] Mark Isbitts. 2017. Preventing Health Care Fraud with Big Data and Analytics. Online. (2017). http://www.lexisnexis.com/risk/insights/health-care-fraud-layered-approach.aspx

[6] Vinil Menon and Parikshi Sheth. 2016. Big Data Analytics Can Be a Game Changer for Healthcare Fraud, Waste, and Abuse. Online. (04 2016). https://www.hfma.org/Content.aspx?id=47523

[7] United States Department of Health and Human Services. 2015. 2015 CMS Statistics. Online. (12 2015). https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMS-Statistics-Reference-Booklet/Downloads/2015CMSStatistics.pdf

[8] United States Department of Health and Human Services. 2016. FY 2016 Agency Financial Report. Online. (11 2016). https://www.hhs.gov/sites/default/files/fy-2016-hhs-agency-financial-report.pdf

[9] United States Department of Health and Human Services. 2017. Combating Medicare Parts C and D Fraud, Waste, and Abuse Web-Based Training Course. Online. (01 2017). https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/CombMedCandDFWAdownload.pdf

[10] Inc. SAS Institute. 2017. Health Insurance Get full coverage for your big data challenges. Online. (2017). https://www.sas.com/en_us/industry/health-insurance.html

[11] Inc. SAS Institute. 2017. SAS Fraud Framework for Health Care. Online. (2017). https://www.sas.com/en_us/software/fraud-framework-for-health-care.html

[12] Healthcare Fraud Shield. 2017. FWAShield A Fully Integrated Fraud, Waste and Abuse Software System. Online. (2017). http://www.hcfraudshield.com/fwashield.aspx

[13] Rodger Smith. 2016. Using Big Data in the Hunt for Healthcare Fraud, Waste, and Abuse Payers must leverage all the big data analytics tools at their disposal to hunt down healthcare fraud, waste, and abuse. Online. (04 2016). https://revcycleintelligence.com/news/using-big-data-in-the-hunt-for-healthcare-fraud-waste-and-abuse

[14] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, and Jos van Hillegersberg. 2013. Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection. *Procedia Technology* 9, Supplement C (2013), 1252 – 1264. https://doi.org/10.1016/j.protcy.2013.12.140 CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

Big Data Applications in Improving Patient Care

Janaki Mudvari Khatiwada
Indiana University
107 S Indiana Ave
Bloomington, IN 47408, USA
jmudvari@iu.edu

ABSTRACT

Big data and its applications in providing the best service outcome to the patients is a new trend. Patient care is the main objective of healthcare organizations. Getting best possible care in terms of costs and service outcome are patients' expectations. How service providers in health-care industries are using big volume of health related data that are generated when patients provide information about their family history, medical history, food and exercise habit or results from clinical tests? Besides health related data, patients are frequently requested to fill out survey about their overall experiences when they get services. They are even asked to give any recommendations to improve their service. There is tremendous use of this type of information in improving patients care services.

KEYWORDS

i523, hid330, Big Data, Health Care, Patient care, Electronic health records

1 INTRODUCTION

Health care is one of the service sectors where service providers claim to have provided consumers with the best experiences possible, whereas consumers are always researching for the best care facilities that they could possibly get which might save time and money and help them have a quality of life. Health service providers collect high volume of information from the consumers every time when they visit the facilities. The volume of health related informations generated in a high velocity is what consist of big data in health care sector. These informations besides clinical records can be anything related to a person. Such as person's ethnic background, exercise routine and the time he or she spends on it on a weekly or daily basis, general daily meal the person intakes, records on wearable health devices and monitors. In today's world, big data has become very impactful in policy making, solving problems and making prediction on whole range of areas. Healthcare industry has become one of the most important sectors to make of use of big data. Big data provides helpful insights for prevention, prediction, diagnosis and identification of best treatment option among all, on the basis of insurance plan a person has. Clinical practitioners acquire, share, compare and analyze big data trend to make their medical diagnosis, treatment recommendation, and prognosis. "A richer set of near real time information can greatly help physicians determine the best course of action for their patients, discover new treatment options, and potentially save lives" [3]. Consumers on the other hand use service provider's web portals to have an insight of available facilities and physicians. Often time we look for ratings and reviews and pick the facility and physician based on those reviews. Big data applications in health care for the purpose of improving patient care is wide; for example, disease prevention and management, health education, research and development, prognosis information sharing, public and individual health management, medical optimization. "A goal of modern healthcare systems is to provide optimal health care through the meaningful use of health information technology in order to improve health care quality and coordination, so that outcomes are consistent with current professional knowledge" [7].

2 APPLICATIONS

Health data are stored as electronic medical records (EMRs), electronic health records (EHRs) or any unstructured records, which are analyzed and shared among clinicians. "These data are near real time data. The EHR, being adopted in many countries, offers a source of data the depth of which is almost inconceivable. About 500 petabytes of data was generated by the EHR in 2012, and by 2020, the data will reach 25,000 petabytes" [2]. One of the trending examples of application of big data in tackling opioid crisis in US. Data professionals at Blue Cross Blue Shield have initiated collaboration with big data experts at Fuzzy Logix to deal with the situation. Data analysts at Fuzzy Logix have used insurance and pharmacy data and identified 742 risk factors which accurately predict people at risk for abusing opiods [5]. "ZEO, Inc. is analyzing over a million nights of data to help consumers improve their sleep" [2]. In general, applications of big data in health care for improving patient care can be categorized into following categories: Prevention, Prediction, Diagnosis, Disease Management and Research and Development.

2.1 Prediction

Analysis of available health records help make prediction which ultimately benefits general population. Making predictions is one of the most useful applications of big data. Researchers use analysis of medical records to make prediction of patients at risk to a disease. "The United States National Institutes of Health has a project known as Pillbox, in which big data are used through the National Library of Medicine" [4]. As a way to monitor flu outbreak, researchers at John Hopkins University (Baltimore, Maryland) created "twitter surveillance system" during flu season year of 2012, 2013 [1]. Similarly, "The Seton Healthcare Family (Austin, TX, USA) and IBM Joint Development Program have done a collaborative work of tracking and analyzing patients' medical information and predicted outcomes of two million patients per year" [10]. Prediction models developed by data analysis are useful in understanding epidemics and finding the best approach to deal with it. This helps in population health management. "Optum Labs has collected EHRs of

over 30 million patients to create a database for predictive analytics tools that will help doctors make big data informed decisions to improve patients treatment" [7]. Use of available database and newly developed predictive models by service providers might help patients and caretakers save time and money. For example, "Parkland Health and Hospital System in Dallas, has generated a valid EHR based algorithm to predict readmission risk in patients with heart failure" [8]. "Those who are at high risk for readmission are provided evidence based interventions, including education and follow-up telephone support within two days of discharge to ensure medication adherence, an outpatient follow up appointment within seven days, and a non-urgent primary care appointment" [8].

2.2 Prevention

The mantra, "Prevention is always better than cure" is what everybody wants to follow. Till now physicians have been studying the general pattern of people's lifestyle and make a recommendations on keeping as it is or make a change to prevent their patients from any health problems. Big data help them identify vulnerable population and raise awareness. For example, physicians recommend general public to watch their weight in order to prevent them from diabetes and heart disease. Another such example is, physicians have identified certain population of certain race are more prone to skin cancer when exposed to sun's ultraviolet rays while other race is more prone to have breast cancer. So, they raise awareness and make needed recommendations accordingly. This in totality help make general public's life better and help them live longer and healthy life. Now we have smart-phones and wearables to track our fitness in general, which generate huge volume of data at a high velocity. In the near future, physicians might be using these data to have an understanding of any potential problem and prepare them for necessary remedies. Collaborations between healthcare and data analytics professionals may be fruitful for predicting future problems and identifying the best available prevention approach [6]. "One recently formed example of such a partnership is the Pittsburgh Health Data Alliance – which aims to take data from various sources (such as medical and insurance records, wearable sensors, genetic data and even social media use) to draw a comprehensive picture of the patient as an individual, in order to offer a customized healthcare package" [6]. "100Plus, a personalized health prediction institute is utilizing uses public and private health and habits data to motivate consumers to take small healthy steps to change daily habits through a mobile application" [2]. This application help consumers focus on preventative measures towards their future health. This is a commendable example of use of big data for improving patient care.

2.3 Diagnosis

Early diagnosis of a disease helps in early intervention of disease management thereby saving lives and reducing costs. Prediction models developed by data analytics researchers by using big data help in early diagnosis. "Predictive modeling over data derived from electronic health records (EHRs) is being used for early diagnosis and is reducing mortality rates from problems such as congestive heart failure and sepsis" [7].

2.4 Disease Management

Early diagnosis might help patients' disease management less complicated because of early interventions. Wearable sensors, monitors and other smart devices help both caregivers and patients to keep track of any changes in factors that is affecting their health. "Processing real time events with machine learning algorithms can provide physicians with insights to help them make lifesaving decisions and allow for effective interventions" [7]. Data about an individual and community reveal informations to physicians and their patients which is helpful in determining appropriate treatment option [9].

2.5 Research and Development

Big data from past help physicians identify general variables responsible for illnesses. After identifying general trend, they can make precise recommendation to their patients and thereby help them have a quality of life and save them costs. Research and development is one of the important applications of big data and analytics that helps in finding new tools, more effective medications, drugs and treatment regimen. Researchers and pharmaceuticals use available flu data to predict a model for the next flu season and develop new flu shot necessary to deal with the outbreak. "Data sharing arrangements between the pharmaceutical giants has led to breakthroughs such as the discovery that desipramine, commonly used as an anti–depressant, has potential uses in curing types of lung cancer" [6]. Big data helps Pharmaceuticals reduce cost of research and therefore lowers drugs cost which benefits patients. Data from clinical trials and patients records help identify adverse effects of a drug.

3 CHALLENGES

While big healthcare data and applications and analytics provides a huge opportunity in improving patient care, it equally comes with some challenges. Privacy and security of personal information is one of the biggest challenges. "In February of 2015, the largest ever healthcare related data theft took place, when hackers stole records relating to 80 million patients from Anthem, the second largest US health insurer" [6]. Since healthcare data are large in volume and are in variety of forms; structured or unstructured, managing this big data of such variety is a challenge. Transforming big volume of unstructured data data which comes in such a velocity, into structured version is another challenge. Data sharing between institutions is another challenge. Maintaining privacy of people's records can be a huge liability for the organizations involving in information sharing.

4 CONCLUSION

While big data in healthcare has some challenges, it has been using for variety of purposes. Above discussed use cases in prediction, prevention, diagnosis, disease management and research and development show the significance of big data and analytics in improving patient care. There is an increasing trend in making use of patients' clinical records for analytics. Going through literatures indicate that use of big data in improving patient care is in the beginning phase and have tremendous potential in the future. Information technology has provided consumers with variety of

wearables making people conscious about their health. In near future physicians might make use of data from the wearables to have an understanding of patients health. Health insurance companies might use big streaming data from wearables to provide incentive such as lowering insurance premium or rewards point to people who are consistent in exercising.

ACKNOWLEDGMENTS

I would like to thank prof.Gregory Von Laszweski and teaching assistants who helped me throughout my writing.

REFERENCES

[1] David A.Broniatowski, Michael J. Paul, and Mark Dredze. 2013. fiNational and Local Influenza Surveillance through Twitter. Online Journal. *PLoS One* 8, 12 (Dec 2013), e83672. https://doi.org/10.1371/journal.pone.0083672

[2] B Feldman, E Martin, and T Skotnes. 2012. *Big data in healthcare hype and hope.* Technical Report. GHDonline. https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf

[3] Hewlett Packard. 2014. *Big Data and healthcare.* Business White Paper. HP. http://h20195.www2.hpe.com/V4/getpdf.aspx/4aa5-2847enw

[4] Rae Jesano. 2010. Free Drug Information Sources on the Web: Government Sites. Online Journal. *Journal of Hospital Librarianship* 10, 2 (Apr 2010), 145–151. https://doi.org/10.1080/15323261003681554 arXiv:http://dx.doi.org/10.1080/15323261003681554

[5] Mona Lebied. 2017. 9 Examples of Big Data Analytics in Healthcare That Can Save People. Web Page. (2017). http://www.datapine.com/blog/big-data-examples-in-healthcare/

[6] Bernard Marr. 2015. *How Big Data is Changing Healthcare.* Blog. Forbes. https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/2/#667047cf140c

[7] Carol McDonald. 2017. *5 Big Data Trends in Healthcare for 2017.* Web Page. MAPR Data Technologies. https://mapr.com/blog/5-big-data-trends-healthcare-2017/

[8] Ravi B. Parikh, Ziad Obermeyer, and David Westfall Bates. 2016. *Making Predictive Analytics a Routine Part of Patient Care.* Web Page. Harvard Business Review. https://hbr.org/2016/04/making-predictive-analytics-a-routine-part-of-patient-care

[9] Wullianallur Raghupathi and Viju Raghupathi 3. PMC. Web. 6 Oct. 2017. 2014. Big Data Analytics in Healthcare: Promise and Potential. *Health Information Science and Systems* 2, 3 (February 2014), 1–10. https://doi.org/10.1186/2047-2501-2-3

[10] S.Feldman, J. Hanover, C. Burghard, and D. Schubmehl. 2012. *Unlocking the power of unstructured data.* White Paper HI235064. IDC Health Insights, 5 Speen Street Framingham, MA 01701 USA. http://uhcjsc.com/pdf/Unlocking%20the%20Power%20of%20Unstructured%20Data.pdf

Big Data Applications in Population Health Management

Tyler Peterson
Indiana University - School of Informatics, Computing, and Engineering
711 N. Park Avenue
Bloomington, Indiana 47408
typeter@iu.edu

ABSTRACT

Healthcare providers are experiencing pressure to reduce costs while also delivering increasingly high quality care. External forces in the form of alternative reimbursement models and programs offering both incentives and penalties have spurred healthcare organizations to find opportunities for increasing the value of their work. Given the complexity of the healthcare system, the opportunities are vast in number. Preventing hospital admissions, management of chronic conditions and the early detection of potentially deadly conditions are a few of the major initiatives. The shared attribute of these three opportunities is that the solutions are often most effective when directed towards people who are going about their day-to-day lives in the community, as opposed to those who are currently confined to hospital beds or in an exam room. Through these programs, providers are compelled to proactively reach out to all people who make up the population they serve. To understand the needs of a population, healthcare providers must embrace big data. Novel analysis and presentation of massive, heterogenous datasets is essential. Big data applications and analytical tools are appearing at the bed side, in exam rooms and on patients themselves as providers seek to harness the power of big data.

KEYWORDS

i523, hid331, Population Health, Predictive Analytics, Electronic Medical Records, EMRs

1 INTRODUCTION

The United States spent $3.2 trillion on healthcare in 2015, 5.8 percent higher than the previous year [8]. This amounts to nearly $10,000 per person living in America [8]. Chronic diseases such as diabetes, cancer and cardiovascular disease contribute to 70 percent of US deaths and incur 75 percent of healthcare expenditures [9]. Despite these high costs, the US lags behind other countries in quality [7]. In other words, we are paying more for less. Healthcare providers are looking to use big data to aid their efforts of reducing costs and increasing the quality of care. The system has a massive amount of data at its disposal, and the rate at which it is accumulating is increasing. In 2011, the US health system possessed 150 exabytes, and the total will soon be counted in zettabytes [14].

The proposed solutions for achieving less expensive, higher quality care are numerous and varied. Several proposals seek to alter the behavior of healthcare providers that have grown accustomed to the fee-for-service reimbursement model. Traditionally, providers are reimbursed for services rendered. Redundant tests, unnecessarily readmitted patients, and frequent emergency room visits generate a profit. This is a financial arrangement that rewards volume without consideration to value [11]. It directs attention to patients who can be immediately provided with services, namely patients who are either in the hospital or at an outpatient appointment. This fee-for-service reimbursement model does not incentivize healthcare providers to look beyond its immediate customers and take the initiative to provide services to the community at-large.

Alternative payment models are engineered with the intention of changing that behavior. The Medicare Shared Savings Program (MSSP) is a type of Accountable Care Organization (ACO) that provides a framework for healthcare organizations to have a level of accountability for the quality, cost and patient experience of an assigned population. MSSP participants choose one of four financial risk arrangements. One track lets providers avoid any penalty, or risk, in the case that they do not lower their Medicare expenditure growth. The other three tracks offer increasingly higher risks and rewards [4]. In order to receive a share of any cost savings, ACOs must demonstrate the delivery of high quality care by reporting the organization's performance on quality measures that can be categorized into four domains - patient experience, care coordination, preventative health and at-risk population management [2].

Other initiatives penalize providers for delivering sub-standard care. In 2016, the US government penalized 2,597 hospitals for excessive 30-day hospital readmissions [15]. If patients are initially admitted with a heart attack, heart failure, pneumonia, chronic lung disease or for a hip/knee replacement procedure and are subsequently readmitted to a hospital within a month of the initial stay, a hospital is considered responsible. The penalties for those readmissions amounted to $528 million nationally in 2016, $108 million higher than 2015 [15].

Entire communities stand to benefit from these programs. Providers must proactively engage with patients who are going about their day-to-day lives while also providing high quality care within their hospitals and at their clinics. To understand how to approach each individual that makes up their community, healthcare providers must harness the power of big data. Mitigating a hospital admission before it occurs or discovering and treating a condition before it worsens requires methodical data collection, pinpointed data analysis, and deliberate, compassionate execution of proactive healthcare delivery. Big data applications and analytical tools are essential for accomplishing those demands.

2 BIG DATA IN POPULATION HEALTH

Big data applications and analytics are well-suited for approaching the issues and opportunities described above because while health care is often described in macro terms, the meaningful interactions happen at the micro level. High-level, aggregated datasets may describe a population, but don't provide the necessary depth for understanding how one patient's needs and circumstances are unique from all the rest. There are various types of data that help illustrate

a patient's overall health picture. Medical claim data includes diagnoses, procedures, dates, cost and points of care. Electronic medical records (EMRs) also includes diagnoses, procedures and dates, while also cataloging lab values, free-text notes, images and medication lists. Clinical trial data, patient satisfaction survey data, genomics data and medical device data also contributes to the massive amount of healthcare data available to providers [10].

This data can be used to meet the nuanced demands of alternative reimbursement models and avoid readmission-related penalties. These programs have attributes that lend themselves to big data applications. MSSP participants need to attest to 31 measures in 2017. These measures, for example, address diabetics with poor hemoglobin A1c control, all-cause, unplanned admissions for heart failure patients, use of imaging studies for low back pain, and patients' perceived quality of communication with providers [3]. Data is essential for identifying patients who fall within the scope of each metric, determining which patients have already met the goal of the measure, and engineering processes to help make providers aware of the patients who have yet to receive the recommended intervention. For example, a healthcare organization must identify their diabetic patients (typically with ICD-10 diagnosis codes), determine which of those patients had their hemoglobin A1c tested within the measurement period, and had a lab result within the accepted value range. Patients may fail the measure in one of two ways: a patient either has not been tested within the measurement period or the patient has a lab value outside of the acceptable range. Patients in the former category should be contacted by the provider and have an appointment scheduled to have the lab drawn. Patients in the latter category should be treated in a manner that brings the hemoglobin A1c within the acceptable range and followed closely by a care team.

To avoid penalties associated with readmissions, some healthcare providers are employing advanced techniques, such as machine learning, to identify patients who are at high risk of being readmitted within 30 days of the initial hospital stay. Mount Sinai Health System in New York, NY, developed a predictive model to evaluate heart failure patients for risk of readmission. The model building began by analyzing 4,205 attributes, including 1,763 diagnosis codes, 1,028 medications, 846 laboratory measurements, 564 surgical procedures, and 4 types of vital signs [13]. Mount Sinai concluded that their model featured in the research study outperformed the previous models used to assess their heart failure patients, while conceding that the model needs to be updated and recalibrated with several years of data from several different hospital sites [13]. In other words, even more data is needed. In the meantime, this model can still be used to analyze each heart failure patient prior to discharge for the likelihood of readmission. Care teams can then dedicate extra resources to especially high-risk patients.

Wearable technology has also infiltrated the healthcare space, especially devices that can remotely and wirelessly monitor patients' vitals and symptoms. The data feeds can be used by providers to assess the effectiveness of (and adherence to) medications, observe lifestyle habits, or recommend that a patient schedule a follow-up appointment or go to an emergency room [1]. There are hundreds of thousands of mobile health apps available in app stores, and more than half of these are geared for patients with chronic diseases [1]. This technology has appeared in clinical trials as well. A study determined that patients with type 2 diabetes who monitored blood glucose with an app achieved greater reduction in hemoglobin A1c results compared to patients who did not use an app [6].

3 INFRASTRUCTURE

The infrastructure needed to support these efforts is complex. A cornerstone of enabling big data analysis in healthcare is EMR software. EMRs replace the paper chart as the location for all details related to patient care. These information systems gather a wide variety of information, including patient encounters, lab results, medications, diagnoses, and procedures, as well as demographic and socioeconomic information, among many other data elements. Providers may also add notes by typing or through dictation software. The information is stored in data warehouses that can be queried and analyzed in a way unimaginable in the era of paper charts. EMRs can also be programmed to remind or notify a provider that, for example, a patient meets the criteria for the MSSP colonoscopy screening measure and has not had a colonoscopy in ten years, so an appointment for the procedure should be scheduled. As of May 2016, 96 percent of non-Federal acute care hospitals had adopted a certified EMR [5].

Health care data is growing in such a way that it benefits from big data applications such as Hadoop and MapReduce, which create a framework capable of handling massive amounts of structured data, such as discrete lab values and diagnosis codes, and unstructured data, such as physician notes. Hadoop breaks the large datasets into smaller subsets, MapReduce processes those subsets independently and in parallel, and the processed subsets are combined into a final result [12].

Data visualization tools are also essential for communiciating key messages in data. Tools such as Tableau and Qlikview, and open source code libraries such as Plotly and Bokeh (written for Python), allow savvy users to present large, complex data sets in visually compelling ways to quickly communicate important ideas. Dashboards can promote exploratory data analysis, and can engage even those who are not technical through easy to use point-and-click user interfaces.

4 CONCLUSION

Big data applications are capable of turning data into insights, and this is critical for aiding healthcare providers in their efforts to evolve the way they practice medicine. EMRs will continue to amass vast amounts of information about patients and their unique characteristics and needs. Programs and policies will continue to foster the mindset that healthcare providers must actively consider all individuals who constitute the population they service, not just the patients actively in a hospital or present in a clinic. Big data applications will continue to be engineered to deliver the right information to the right provider. These tools will promote the most beneficial action for each individual patient.

ACKNOWLEDGMENTS

The author would like to thank Professor Gregor von Laszewski and his teaching assistant for their help with Github, Latex and JabRef.

REFERENCES

[1] Linda Brookes. 2017. Can Technology Transform Chronic Disease Management? Online. (April 2017). https://www.medicalnewstoday.com/articles/317016.php

[2] Centers for Medicare and Medicaid Services. 2017. Accountable Care Organization 2017 Quality Measure Narrative Specifications. Online. (January 2017). https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/2017-Reporting-Year-Narrative-Specifications.pdf

[3] Center for Medicare and Medicaid Services. 2017. Medicare Share Savings Program Quality Measure Benchmarks for 2016 and 2017 Reporting Years. Online. (December 2017). https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/MSSP-QM-Benchmarks-2016.pdf

[4] Centers for Medicare and Medicaid Services. 2017. New Accountable Care Organization Model Opportunity: Medicare ACO Track 1+ Model. Online. (July 2017). https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/New-Accountable-Care-Organization-Model-Opportunity-Fact-Sheet.pdf

[5] JaWanna Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. 2016. Adoption of Electronic Health Record Systems Among U.S. Non-Federal Acute Care Hospitals: 2008-2015. Online. (May 2016). https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php

[6] Can Hou, Ben Carter, Jonathan Hewitt, Trevor Francisa, and Sharon Mayor. 2016. Do Mobile Phone Applications Improve Glycemic Control (HbA1c) in the Self-management of Diabetes? A Systematic Review, Meta-analysis, and GRADE of 14 Randomized Trials. *Diabetes Care* 39, 11 (November 2016), 2089–2095. http://care.diabetesjournals.org/content/39/11/2089

[7] Rabah Kamal and Cynthia Cox. 2017. U.S. Health System is Performing Better, Though Still Lagging Behind Other Countries. Online. (May 2017). https://www.healthsystemtracker.org/brief/u-s-health-system-performing-better-though-still-lagging-behind-countries/#item-start

[8] Anne B. Martin, Micah Hartman, Benjamin Washington, Aaron Catlin, and National Health Expenditure Accounts Team. 2017. National Health Spending: Faster Growth In 2015 As Coverage Expands And Utilization Increases. *Health Affairs* 36, No. 1 (2017), 166–176. http://content.healthaffairs.org/content/36/1/166

[9] Farshad Fani Marvasti and Randall S. Stafford. 2012. From Sick Care to Health Care - Reengineering Prevention Into The U.S. System. *The New England Journal of Medicine* 367, 10 (September 2012), 889–891. http://www.nejm.org/doi/full/10.1056/NEJMp1206230#t=article

[10] Carol McDonald. 2016. How Big Data Is Reducing Costs And Improving Outcomes In Health Care. Online. (June 2016). https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/

[11] Harold D. Miller. 2009. From Volume To Value: Better Ways To Pay For Health Care. *Health Affairs* 28, 5 (2009), 1418–1428. http://content.healthaffairs.org/content/28/5/1418.abstract

[12] Elizabeth O'Dowd. 2016. How Hadoop Supports Healthcare Data Analytics Infrastructure. Online. (October 2016). https://hitinfrastructure.com/news/how-hadoop-supports-healthcare-data-analytics-infrastructure

[13] Pacific Symposium on Biocomputing (Ed.). 2016. *Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort.* Vol. 22. Pacific Symposium on Biocomputing. https://www.ncbi.nlm.nih.gov/pubmed/27896982

[14] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big Data Analytics In Healthcare: Promise And Potential. *Health Information Science and Systems* 2, 3 (2014), 1–10. https://www.biomedcentral.com/track/pdf/10.1186/2047-2501-2-3?site=hissjournal.biomedcentral.com

[15] Jordan Rau. 2016. Medicare's Readmission Penalties Hit New High. Online. (August 2016). https://khn.org/news/more-than-half-of-hospitals-to-be-penalized-for-excess-readmissions/amp/

Big Data Analytics, Data Mining, and Public Health Informatics: Using Data Mining of Social Media to Track Epidemics

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

ABSTRACT

Data mining of internet search queries and social media for influenza related keywords has been used to track seasonal influenza and correlates highly with official reports of 'infuenza-like-illness' (ILI). Efforts to monitor epidemics using big data analytics can provide early detection that supplements existing systems of disease surveillance. A review of the literature shows that data extracted from social media has applications for public health informatics. Prediction models based on social media work best in areas with a high degree of internet access.

KEYWORDS

i523, HID335, Data Mining, Social Media, Public Health Informatics

1 INTRODUCTION

In the information age, *Big Data* offers great promise to fuel innovation, generate new revenue streams, and transform society [10]. Can the potential of big data be harnessed for the greater good, to prevent disease and improve health? Seasonal influenza epidemics are a major public health concern, resulting each year in an estimated 250,000 to 500,000 deaths worldwide [16]. This paper explores big data in public health informatics, specifically reviewing research on data mining to track epidemics and the spread of contagious disease [11]. Can these approaches be extended to monitor other epidemics such as the opioid crisis in North America? [20] Epidemic spreading is a complex phenomenon based on contact networks between individuals and distributed by transportation networks [5]. Some questions remain as to whether prediction models based on social networking platforms can be generalized to other epidemics at future points in time. Limitations of using social media data to predict epidemics are discussed.

1.1 Public Health Informatics

The field of Health Informatics is generating huge amounts of data at a rapid pace, from MRI imaging data, electronic medical records (EMRs), clinical research data, to population-level data. This review focuses on population data from search queries and social media to provide insights about epidemics and pandemics [11, 12]. Big data is an ambiguous term that lacks a single unified definition, but is often described in terms of *Volume, Velocity, Variety, Veracity, and Value* [6]. Trying to track an epidemic in real-time from multitudes of incoming web searches and posts involves a high volume of data coming in at high velocity [14, 18]. In order to be of any use, diverse and often messy raw data has to be sifted through and effectively organized for further analysis. The issue of Veracity raises the questions of how reliable social media data are for predicting real life events. What is the relationship between social media data to biological events such as the spreading of contagion and disease? The question of Value evaluates the quality of the data as it pertains to intended outcomes, such as limiting the spread of contagion and disease prevention. There are legitimate concerns about the quality of data obtained from the internet; however, the literature suggests that mining information from social media can produce valuable data. An important challenge for making sense of big data is developing analytic tools adequate to handle large volumes of data in real time.

1.2 Data Mining Social Media

Health Informatics research is considered from two levels: where the data is collected, and the research questions being addressed. Research on social media can yield data on a range of issues related to public health, including: spatiotemporal information of disease outbreaks, real-time tracking of infectious diseases, global distributions of various diseases, and search queries on medical questions that people might have [12]. The questions of interest in the current review are: *Can search query data be used to accurately track epidemics in real-time?* and, *can Twitter data be used to monitor epidemics across different regions?* The general idea is that increasing search query or social media activity is associated with an increasing interest in a given health topic. A limitation of social media data is that, although it has high Volume, Velocity, and Variety, it can be unreliable, resulting in both low Veracity and Value [11, 15]. A review of the literature shows how useful data can be extracted by data mining and analytic techniques.

1.3 Using Search Queries to Track Epidemics

1.3.1 *Tracking Epidemics Using Google Search Terms in the U.S..*

Seasonal influenza is an acute viral infection that spreads easily from person to person, circulates across regions, affecting people of every age. Traditional flu monitoring estimates from the U.S. Center of Disease Control and Prevention (CDC) based on physician reports of patients with "*influenza-like illness*" (ILI) are released weekly [4], but generally with a one to two week delay. In an effort to improve on early detection of season influenza, a team of researchers developed an automated method to analyze Google search queries to track ILI terms from historical logs between 2003 and 2008, using 50 million most popular searches, and CDC historical data [7]. The *Google Flu Trends* (GFT) model [8] sought to find the probability that a given search query is related to an ILI of a patient visiting a physician in the same region. GFT used a feature selection method to narrow the 50 million most popular search queries, aggregrated from historical, down to 45. These top 45 search queries yielded the highest estimates during cross validation and were connected with influenza symptoms, complications, remedies, consistent with searches by individuals with influenza. Estimates of the current

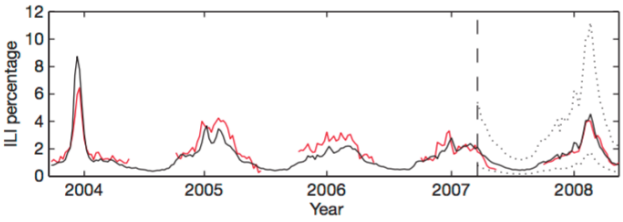

Figure 1: Comparison of GFT model estimates (black) for mid-Atlantic region of U.S. (NY, NJ, PA) against CDC-reported ILI percentages (red) between 2004 to 2008 [7]

level of weekly influenza were based on the correlation of the relative frequency of search queries and the percentage of physician visits with patients presenting influenza-like symptoms. The GFT model was trained on 128 points of the mid-Atlantic region of the U.S. (e.g., New York, New Jersey, Pennsylvania) between 2004 to 2007 with a correlation of 0.85, and validated on 42 points between 2007 to 2008, with a correlation of 0.96 (see Figure 1). The final model, for all regions in the U.S., generated correlation estimates ranging from 0.92 to 0.99 over 42 points. Thus, analyzing high volume Google search queries estimated ILI percentages across several regions in the U.S. about 1 to 2 weeks earlier than official CDC ILI reports. Such efforts at early detection can help physicians and health care professional anticipate and prepare for the outbreak of influenza epidemics and pandemics.

1.3.2 Tracking H1N1 Epidemic Using Baidu Queries in China. Researchers in China monitored influenza activity by comparing internet search query data from *Baidu* (https://www.baidu.com) to influenza case counts from the Chinese Ministry of Health (MOH) between 2009 to 2012 during the H1N1 epidemic [23]. The study consisted of four parts: (i) Selecting keyword terms related to influenza, (ii) Filtering keywords unrelated to flu epidemics, (iii) Defining weights and composite search index, and (iv) Fitting a regression model with keyword index to influenza case data. In the process of filtering, only 40 of 94 keywords were correlated with the case data, and only 8 of these 40 keywords were used as the optimal set in the composite search index. As expected, the search index captured seasonal variation of influenza epidemics in the Winter and Spring, indicating a good predictor for tracking influenza activity in China (see Figure 2). The regression model accounted for 95 percent of the variability in influenza case data (ICD), and the model was validated for a test period in 2012. The mean absolute percent error rate of prediction over an eight month period in 2012 was 10.6 (see Table 1). This research yields additional evidence that novel approaches using big data can provide early indicators of epidemic activity that supplement official public health information sources, rather than replacing them. A limitation acknowledged by the authors is the relatively small initial number of keyword search terms used compared to the Google Flu Trends (GFT) project [7]. Another limitation of using search query data is that, although the keywords selected in this model performed well at capturing temporal trends in the H1N1 epidemic, the same keywords may not reflect the trend of an influenza epidemic at a future time. The authors also noted

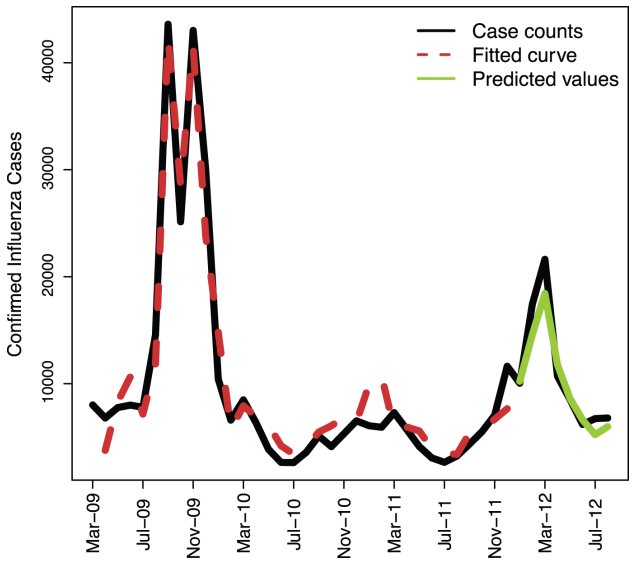

Figure 2: Plot of influenza cases, fitted values and prediction based on model [23]

Table 1: Predicted values, errors, and mean absolute percent error of prediction based on Baidu search queries in China for eight consecutive months (January to August 2012) [23]

Month	Actual values	Predicted Values	Absolute Error	Percent absolute error
01-2012	10045	10230	184	1.8
02-2012	17421	14578	2843	16.3
03-2012	21652	18429	3196	14.8
04-2012	10707	11785	1078	10.1
05-2012	8520	8618	98	1.2
06-2012	6195	6621	426	6.9
07-2012	6738	5240	1498	22.2
08-2012	6793	5983	810	11.9

the lack of internet access in rural areas, which underscores the fact that effective tracking of epidemics based on search queries relies on internet access. Furthermore, caution should be used in evaluating correlational data, as causation cannot be inferred from correlation.

1.4 Using *Twitter* API Data to Track Epidemics

Twitter is a free online social networking and micro-blogging service, where users can send and read messages of 140 characters (i.e., "*tweets*"). As of 2017, Twitter has more than 320 million monthly active users (67 million in U.S.), with an estimated 500 million tweets posted per day (https://about.twitter.com). Twitter users share their perspectives and reactions on a wide range of topics, approximately 80 percent from handheld mobile devices, acting as "sensors" of events in real time [1]. The Twitter stream provides a rich data source for tracking or forecasting general sentiment, political attitudes, linguistic variation, detecting earthquakes, and disease surveillance. The large volume of users provides a high likelihood that ILI epidemic information is posted; however, Twitter

post data is noisy and perhaps unreliable insofar as it can be difficult to differentiate posts about the flu based on instances of concerned awareness (*"I am worried about the swine flu epidemic!"*), versus actual infection (*"Robbie might have swine flu. I am worried."*)[14]. Despite the noise in Twitter data from much useless chatter, useful information be obtained from mining data in the Twitter stream.

1.4.1 Using Twitter to Track Disease Activity and Public Concern in the U.S. during the H1N1 Pandemic. In a 2011 study, researchers searched through post data from Twitter's streaming API during the H1N1 epidemic (October 2009 to May 2010) across spatiotemporal areas of the U.S. to predict weekly ILI levels [19]. Tweets were sifted according to keywords related to H1N1 (e.g., *"flu", "swine", "influenza"*) and additional terms about vaccines, side effects, and/or vaccine shortages. The first data set consisted of 951,697 tweets containing influenza related keywords from 334,840,972 tweets extracted between April to June 2009 (results were reported as a percentage of observed tweets). These tweets represent just over 1 percent of the sample tweet volume, and this percentage declined rapidly over time as the number of reported H1N1 cases increased. In the U.S. surveillance programs track reported influenza-like illness (ILI) seasonally, from October to May, monitoring the total number of patients seen along with the number with ILIs reported. Quantitative estimates of ILI values based on the Twitter stream were analyzed using support vector regression (SVR) and leave-one-out cross-validation to test model accuracy. Figure 3 shows the weekly ILI values nationwide reported by the CDC (green line) and estimated using a model trained on roughly 1 million influenza-related Tweets (red line) obtained between October, 2009 to May, 2010. The red line shows output from a leave one out cross validation based on SVM estimator. Point estimates of national ILI values produced by the system were good with an average error of 0.28 percent. A regional model, based on significantly fewer tweets, approximated the epidemic curve for CDC region 2 (New York, New Jersey) as reported by the ILI data, but the estimate was less precise with an average error of 0.37 percent. In terms of public interest, Twitter users' interest in antiviral drugs dropped, as official disease reports indicated most influenza cases were relatively mild, even as the number of cases was increasing. In addition, interest in hand hygiene and face masks was associated with public health messages from CDC. A limitation of the study is that only a limited number of search terms and one prediction method was used. An important question is whether the results could be improved using broader search terms and other prediction models.

1.4.2 Twitter Improves Seasonal Influenza Prediction. In a 2012 study, researchers implemented a system using an online social network (OSN) Crawler bot to retrieve tweets by keywords (e.g., *"flu", "H1N1", "swine flu"*), geospatial location, relative keyword frequency, and CDC ILI reports [1]. The *Social Network Enabled Flu Trends* (SNEFT) network continuously monitored tweets and profile details of the Twitter users who commented on flu keywords (starting October 2009), to detect and track the spread of ILI epidemics. The correlation between flu related tweets and ILI was very high between 2009 to 2010 (r=0.98) during the H1N1 outbreak, but the correlation dropped substantially for 2010-2011 (r=0.47) after the epidemic, suggesting that noisy tweets became more prominent as H1N1 was less of an issue. To reduce noise, text classification using

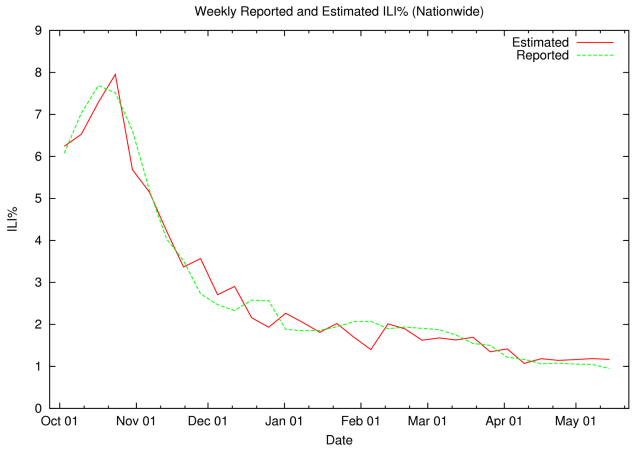

Figure 3: Weekly CDC reported (green) and Twitter estimated (red) ILI percent Nationwide (U.S., 2009 to 2010) [19]

support vector machines (SVM) was trained on a dataset of 25,000 tweets to determine whether a tweet was related to a flu event or not; data cleansing was conducted to remove multiple tweets posted by the same user during a single bout with the same illness. These methods improved the correlation between the Twitter data and ILI rates from the CDC from October 2010 to May 2011 in the U.S. (r=0.89), and Twitter data was correlated with ILI rates across subregions. Figure 4 shows the weekly plot of percentage weighted ILI visits, positively classified Twitter users and predicted ILI rate using CDC and Twitter for 2010 to 2011. The authors reported that Twitter data alone had higher prediction rates toward the beginning and end of the flu season, and during an epidemic; however, the also noted that using previous CDC ILI data offered a better assessment for making flu predictions. In addition, age analysis suggested Twitter data best fit the age groups of 5-24 years and 25-49 years, for most regions in the U.S. The results showed Twitter data can be used to detect and possibly predict ongoing ILI epidemics in real time with relatively low error, up to 1-2 weeks earlier than the CDC reportings. It would be interesting to determine whether these results could be generalized beyond the U.S. and replicated with populations in other countries [23].

1.5 Limitations of Using Search Queries and Social Media Data to Track Epidemics

There is some evidence that influenza forecasting models based on Twitter data performed better than general search query data [18]. Google Flu Trends (GFT) algorithms underestimated ILI in the U.S. at the start of the H1N1 (i.e. *swine flu*) pandemic in 2009 [2], and over-predicted seasonal influenza in January 2013 compared to the CDC ILI by almost double [15] (shown in Figure 5). As described above, there are important limitations in using social media data for predicting epidemics: First, internet access and Twitter usage is not uniform by geographical region. Urban areas have higher density of internet connections than rural areas [23], and coastal regions

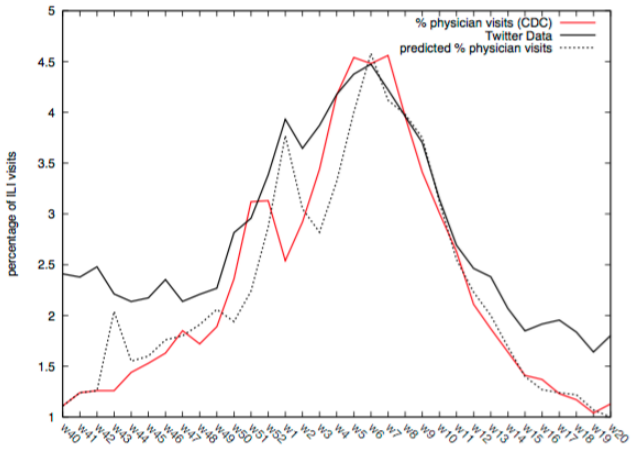

Figure 4: Weekly plot of percentage weighted ILI visits, positively classified Twitter dataset and predicted ILI rate using CDC and Twitter [1]

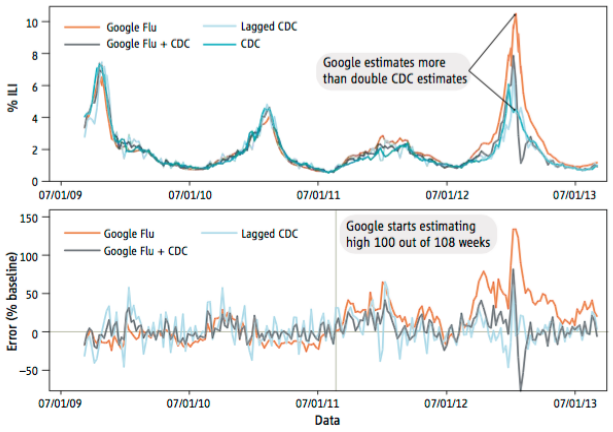

Figure 5: GFT overestimation: GFT overestimated the prevalence of flu in 2012-2013 season and overshot the actual level in 2011-2012 by more than 50 percent [15]

of the U.S. (CA, NY) produced more tweets per person than Midwestern U.S. states (or Europe) [1]. Thus, performance of seasonal influenza predictions models may be best applied to regions with high internet access and where tweets are more frequent. Second, exact demographic information about the Twitter population is not easy to estimate (or unknown) and the demographic of internet users does not represent characteristics of the general population. Third, though promising, the results of this research are based on correlations between often noisy internet search queries or Twitter posts and physician reports of ILI compiled by official governmental sources. Caution should be used in evaluating predictions about serious health concerns such as epidemics or pandemics based on correlational evidence as the data do not support causal inferences.

1.6 The Dynamics of Epidemic Spreading

Can these methods be extended to survey other types of epidemics? The dynamics of epidemic spreading is a complex phenomenon, based on contact networks of person-to-person interaction, indirect exposure, and transmission byways such as the *airline transportation network* (ATN) [5]. Epidemics are quantified in terms of the proportion of the population infected, those yet to be infected, and the rate of transmission [13]. In addition, the structure of the contact network can influence epidemic spreading [17]. For example, in the case of simple contagion, weak ties among acquaintances or infrequent associations provide shortcuts between distant nodes that reduce distance within the network [9] and can facilitate the spread of disease. Furthermore, networks with "small world" properties have many nodes with few connections, but a small number of highly connected nodes that can rapidly transmit contagion throughout the network [22]. Analyzing the correlation between Twitter posts and rate of ILI reports does not capture the complex network structure that underlies disease epidemics and pandemics. It is possible that by analyzing the structure of social media networks, future research may help to identify how points of connection within online networks are associated with the spread of contagion and resulting epidemics [24]. Some epidemics such as the opioid crisis in North America [21] may be amenable to social network modeling as drug usage, dependancy, and addiction is subserved by social networks. The emergence of new technologies, such as wearable biosensors [3] may help improve geospatial mapping of the opioid epidemics and treatment interventions.

2 CONCLUSION

Big data mining of social media has tremendous potential to detect trends and confirm observations based on real time events, providing opportunities to monitor infectious disease on a global level. The research reviewed above shows how search queries and Twitter data about ILI related information provides an early detection signal that can supplement existing epidemic monitoring systems and may help improve public health responses and prevention. As described above, these approaches to tracking disease and predicting epidemics work best in areas with high internet connectivity and are better suited to populations with a high proportion of social media users.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for providing the LaTex template and instructions, many comments, helpful feedback, edits, and fixes. Thanks also to the Assistant Instructors who were also very helpful in this learning process.

REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Liu. 2012. Twitter improves seasonal influenza prediction.. In *International Conference on Health Informatics*. International Conference on Health Informatics, Vilamoura, Algarve, Portugal. DOI: http://dx.doi.org/~bliu/pub/healthinf_2012.pdf

[2] D. Butler. 2013. When Google got flu wrong. *Nature* 494, 7436 (Feb. 2013), 155–156. DOI: http://dx.doi.org/10.1038/494155a

[3] S. Carreiro, D. Smelson, M. Ranney, K. J. Horvath, R. W. Picard, E. D. Boudreaux, R. Hayes, and E. W. Boyer. 2015. Real-Time Mobile Detection of Drug Use with Wearable Biosensors: A Pilot Study. *Journal of Medical Toxicology*. 11, 1 (March 2015), 73–9. DOI: http://dx.doi.org/10.1007/s13181-014-0439-7

[4] C.D.C. 2017. *FluView: Weekly U.S. Influenza Surveillance Report.* Technical Report. U.S. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/weekly/index.htm

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

[6] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. De Laat. 2012. Addressing big data challenges for scientific data infrastructure. In *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, IEEE, Taipei, Taiwan, 614–617. DOI: http://dx.doi.org/10.1109/CloudCom.2012.6427494

[7] J. Ginsberg, M.H. Mohebbi, R. S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 19 (Feb. 2009), 1012–1014. DOI: http://dx.doi.org/10.1038/nature07634

[8] Inc. Google. 2009. Google Flu Trends for the U.S.: Original Estimates (v1). (2009). https://www.google.org/flutrends/about/data/flu/historic/us-historic-v1.txt

[9] M. S. Granovetter. 1973. The strength of weak ties. *Amer. J. Sociology* 78, 6 (May 1973), 1360fi?!1380. DOI: http://dx.doi.org/10.1086/225469

[10] Sunil Gupta. 2015. Big Data: Big Deal or Big Hype? *European Business Review* (May 2015). http://www.europeanbusinessreview.com/big-data-big-deal-or-big-hype/

[11] Simon I Hay, Dylan B George, Catherine L Moyes, and John S Brownstein. 2013. Big data opportunities for global infectious disease surveillance. *PLoS medicine* 10, 4 (2013), e1001413. DOI: http://dx.doi.org/https://doi.org/10.1371/journal.pmed.1001413

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

[13] Herbert W. Hethcote. 2000. The Mathematics of Infectious Disease. *Society for Industrial and Applied Mathematics (SIAM) Review* 42, 4 (2000), 599fi?!653. DOI: http://dx.doi.org/https://doi.org/10.1137/S0036144500371907

[14] Alex Lamb, Michael J Paul, and Mark Dredze. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter.. In *HLT-NAACL*, Association for Computational Linguistics (Ed.). Association for Computational Linguistics, Atlanta, Georgia, 789–795. http://www.aclweb.org/anthology/N/N13/N13-1097.pdf

[15] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6176 (2014), 1203–1205. DOI: http://dx.doi.org/10.1126/science.1248506 arXiv:http://science.sciencemag.org/content/343/6176/1203.full.pdf

[16] World Health Organization. 2016. Influenza (Seasonal),. online. (Nov. 2016). http://www.who.int/mediacentre/factsheets/fs211/en/

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

[18] M. J. Paul, M. Dredze, and D. Broniatowski. 2014. Twitter Improves Influenza Forecasting. *PLOS Currents: Outbreaks* 6, 1 (Oct. 2014). DOI: http://dx.doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[19] A. Signorini, A. M. Segre, and P. M. Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLOS ONE* 6, 5 (May 2011), e19467. DOI: http://dx.doi.org/https://doi.org/10.1371/journal.pone.0019467

[20] M. Smith. 2016. Can social media help prevent opioid abuse? online. (July 2016). DOI: http://dx.doi.org/10.1126/science.aag0661

[21] 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. DOI: http://dx.doi.org/10.1056/NEJMp1402780 arXiv:http://dx.doi.org/10.1056/NEJMp1402780 PMID: 24758595.

[22] D. J. Watts and S .H. Strogatz. 1998. Collective dynamics of fismall-worldfi networks. *Nature* 393, 4 (June 1998), 440–442. http://www.stat.cmu.edu/~fienberg/Stat36-835/WattsStrogatz-Nature-1998.pdf

[23] Qingyu Yuan, Elaine O Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S Brownstein. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS one* 8, 5 (2013), e64323. DOI: http://dx.doi.org/https://doi.org/10.1371/journal.pone.0064323

[24] Yu-Xiao Zhu, Wei Wang, Ming Tang, and Yong-Yeol Ahn. 2017. Social contagions on weighted networks. *Phys. Rev. E* 96 (July 2017), 012306. Issue 1. DOI: http://dx.doi.org/10.1103/PhysRevE.96.012306

Big Data Application for Breast Cancer Treatment

Hady Sylla
Indiana University
Smith Research Center
Bloomington, IN 47408, USA
hsylla@iu.edu

ABSTRACT

This paper is about Big Data application for the treatment of breast cancer.The paper explore contribution that big data had on the analysis of numerous data by researchers

KEYWORDS

HID 339, Big Data, Breast Cancer, Cancer

1 INTRODUCTION

We are living in the time of Big Data, which has encouraged a significant discovery in medicine and will support most, if not all, of the treatment and anticipation progresses yet to come[7]. The International Cancer Genome Consortium. The Cancer Genome Atlas, and by international consortia, e.g., Information inside any one asset can go from a specialty informational index to completely coordinated information created by numerous innovation stages and a large number of patient examples[7]. In the past, scientists were aware of the fact that cancers frequently have extra or missing chromosomes or pieces of chromosomes, which is referred to as aneuploidy

2 BIG DATA APPLICATION FOR BREAST CANCER TREATMENT

There are 3 billion DNA code letters in each human cell and 32 thousand billion cells in the body. So every individual has 96 thousand billion DNA code letters[6]. That is more than ten times the quantity of code letters as there are grains of sand in most of the shorelines on Earth. Besides, a lamentable difference in any of those code letters can begin a sickness or make it impenetrable to medical research[6].

The term breast cancer encompasses many different cancers as no two breast cancers are precisely the same. Researchers utilize genomic technology to describe these cancers fully and develop applications of this knowledge that guide treatment decisions tailored to fit the needs of individual patients[1].But, it was unclear until recently if this characteristic was significant or merely the byproduct of tumor growth[9]. In 2013. The research conducted by geneticist Stephen Elledge identified aneuploidy as the factor that is responsible for driving cancer[9].This discovery was derived using tremendous amounts of cellular data and the ability of computers to aid researchers in sifting through this information[9].

3 BIG DATA TO ADVANCE BREAST CANCER RISK PREDICTION

Researchers supported by Cancer Research have made a 'guide' connecting the state of the city with of breast cancer cells to genes switched on and off, and coordinated it to genuine malady results, as indicated by an examination distributed in Genome Research[5]. Big-data researchers utilize an extensive data set, such as the Cancer Genome Atlas (TCGA), and look for patterns within the data[1]. The goal is to identify mutations, which researchers can target by drug treatment that they personalized to a particular patient's needs[1]. Big data research involves analyzing the data derived from thousands of tumors, which reveal patterns that can improve screenings and diagnosis, as well as guide treatment[1].[4]provide an overview of data resources on cancer-related research. These authors review compendia of data resources, a list of cancer-related data resources, and a biomolecular repository "'Hubs,'" as well as lists of seminal publications and journals on data science[4]. In other words, this review describes where researchers can find breast cancer data and aids determining the range of data types that are available[4].

When a cancer patient is diagnosed, the tumor's genome can be sequenced, and this information can be used to identify drugs that are likely to affect tumor growth (Savage, 2014). Elledge's discovery that aneuploidy is the engine driving cancer growth resulted directly from a computational method developed by his researcher and his colleagues, the Tumor Suppressor and Oncogene Explorer (TSOE)[9]. The TSOE is used to mind large data sets, which include the Cancer Genome Atlas and the Catalogue of Somatic Mutations in Cancer[9]. There were roughly 70 suppressor genes and 50 oncogenes already known, but the development of the TSOE increased these number to approximately 329 and 200 respectively[9]. Analytical data is available on 8,200 tumors, but researchers consider this to be just a start[9].

Furhermore, [11] stated that text-mining using empirical literature makes possible the discovery of new knowledge that will help researchers obtain a better understanding of human diseases, which can then be used to improve the care delivery. These researchers designed and developed a text-mining framework that they refer to as Spark-Text, which utilizes a Big Data infrastructure that includes Apache Spark data streaming and machine learning methods, combined with a Cassandra NoSQL database[11]. The researcher extracted information relevant to several types of cancers, including breast cancer, accessing tens of thousands of articles. The researchers conclude that the potential for mining scientific articles using this Dig Data infrastructure is very high[11]. Furthermore, the SparkText program can be utilized in other areas of biomedical research[11].

Significant problem with the vast data sets relevant to genomic cancer data based on biomarkers concerns developing methods for manipulating this information, which can terabyte level and beyond. Big data infrastructures, such as the one developed by[11]. offers means for utilizing this invaluable data and using this information to

inform screening, diagnosis, and delivery of quality care to patients. Individually, big-data science has led to researchers rethinking how to breast cancer[1].

While mining big data holds the potential of leading to a medical breakthrough, the information is analyzed thoroughly, and it is also necessary to understand the pros and cons of big data analytics[2]. The advantages include the fact that big data focused on correlations, not causality, which means that big data sets have the power to alert researchers to patterns that they did not expect [2].Big data allows healthcare providers to personalized treatment to fit the needs of individual patients. A University of Ontario study of sepsis in premature babies demonstrates how analysis of large data set can provide correlations that lead to clinical actions [2].By employing the data from 1200 data points-per-second, generated from wireless sensors attached to babies, the researchers successfully diagnosed infections 24 hours before fever development and increases in white blood cell count [2].

In another research study distributed by Genome Research, the researchers effectively mapped the state of bosom tumor cells to qualities, and coordinated it to illness results[3]. This guide could help doctors in picking a treatment for patients. The researchers utilized extensive datasets to establish a ink between cell shape and qualities. Generally, they inspected more than 300,000 bosom tumor cells and 28,000 distinct qualities[3]. The investigators found that NF-kappaB is a central protein involved with the network, and could promote proliferation and metastasis of cancer cells. This was linked to cancer stage, and may be used to predict survival outcomes in patients with breast cancer. Through big data approach, the analysts could filter through a huge number of disease cells and qualities to decide their affiliations. This guide could be utilized by doctors later on to decide the treatment alternative that has prompted the most elevated survival rate in patients with comparable malignancies. It could likewise give understanding to both the patient and the doctor about the idea of the ailment.

Furhermore, Madabhushi worked with Shannon C. Agner at Rutgers University and Mark A. Rosen, MD; Sarah Englander; Mitchell D. Schnall, MD; Michael D. Feldman, MD; Paul Zhang, MD; and Carolyn Miles; MD, at the University of Pennsylvania, on the breast cancer study. They broke down MR pictures of bosom injuries from 65 ladies. The specialists filtered through several gigabytes of picture information from every patient to attempt to discover contrasts that recognize the diverse breast cancer subtypes. The researchers scientifically demonstrated the surfaces that show up as the tissues retain differentiate improving color. The model uncovered that progressions over just milliseconds recognized triple-negative from kind sores. The examiners utilized machine learning and example acknowledgment techniques to help in analyze among the three sorts of growths in view of surface changes and other quantitative proof[10]. Madabhushi posited that Today, if a lady or her specialist finds a protuberance, she gets a mammogram and after that a biopsy for atomic examination, which can take two weeks or up to a month. In the event that we can anticipate the malignancy is triple-negative, we can quick track the patient for biopsy and treatment. Particularly in cases with triple-negative malignancy, two to a month spared can be pivotal[10].

A discussion by Clifford Hudis, MD, at the fifteenth St Gallen Breast Cancer Conference stated that The lack of patients taking part in clinical trials makes information extrapolation and application complex, requiring a need to investigate wellbeing innovation arrangements that tap the capability of genuine information[8]. In the United States, just roughly 3% of patients determined to have malignancy are enlisted in clinical trials. Besides, significant difference exists in socioeconomics between members in clinical trials and the all inclusive community. Regardless of electronic record selection, "one impediment to curing malignancy stayed: tolerant information isn't shared," said Hudis. Indeed, wellbeing data got under this demonstration was not interoperable, speaking to a noteworthy obstacle. As a rule, the electronic wellbeing records were fundamentally kept up with the goal that clinicians could enough guard their charging hones for expensive medications and therapeutics upon review. The answer for this issue was a framework called CancerLinQ, noted Hudis, who is the present seat of the huge information activity's governing body. CancerLinQ incorporates quiet information, inside privacy rules, and takes into consideration information mining and sharing. "The basic role of CancerLinQ is to enhance the nature of care and to upgrade results," Hudis stated[8]. By March 2017, almost 2 million records had been joined in the framework from 80 oncology mind settings, which extended from singular practices to huge growth focuses. The framework benefits from the information as of now being entered. In a normal day, around 40% of a clinicians time is currently spent on record entering, as per Hudis. Hudis conclude by stating that traditional research drives us forward yet is restricted by a tight pool of subjects, and the cost and long time expected to build up a result. Later on, huge informational collections may expand and broaden the fantastic confirmation from planned research that incorporates more seasoned patients, comorbidities, simultaneous medicine and numerous other certifiable ramifications of fruitful tumor treatment[8].

4 CONCLUSION

As another field of research, the look for measures that boost predictivity may do much in the method of satisfying the expectations of progressing anticipating results of intrigue. Big data allows the dissecting information from numerous disease sorts that researchers can assess prognostic models and recognize quality changes that prompted tumor formation.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Jill U Adams. 2015. Genetics: big hopes for big data. *Nature* 527, 7578 (2015), S108–S109.
[2] Kent Bottles and Edmon Begoli. 2014. Understanding the pros and cons of big data analytics. *Physician executive* 40, 4 (2014), 6.
[3] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.
[4] Susan E. Clare and Pamela L. Shaw. 2016. Big Data for breast cancer: where to look and what you will find. *NPJ Breast Cancer* 2 (11 2016), 16031. https://search-proquest-com.ezp.waldenulibrary.org/docview/1835693005?accountid=14872 Copyright - Copyright Nature Publishing Group Nov 2016; Last updated - 2016-11-03.
[5] Robert A Hiatt and Barbara K Rimer. 1999. A new strategy for cancer control research. *Cancer Epidemiology and Prevention Biomarkers* 8, 11 (1999), 957–964.

[6] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. 2004. The effects of human population structure on large genetic association studies. *Nature Genetics* 36, 5 (2004), 512 – 517.

[7] Travis B Murdoch and Allan S Detsky. 2013. The inevitable application of big data to health care. *Jama* 309, 13 (2013), 1351–1352.

[8] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* 2, 1 (07 Feb 2014), 3. https://doi.org/10.1186/2047-2501-2-3

[9] Neil Savage. 2014. Bioinformatics: big data versus the big C. *Nature* 509, 7502 (2014), S66–S67.

[10] Samuel Fosso Wamba, Shahriar Akter, Andrew Edwards, Geoffrey Chopin, and Denis Gnanzou. 2015. How fibig datafican make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics* 165 (2015), 234–246.

[11] Zhan Ye, Ahmad P Tafti, Karen Y He, Kai Wang, and Max M He. 2016. Sparktext: Biomedical text mining on big data framework. *PloS one* 11, 9 (2016), e0162721.

Sociological Applications of Big Data

Jeramy Townsley
IUPUI
425 University Ave
Indianapolis, Indiana 46202
jtownsle@indiana.edu

ABSTRACT

The social sciences have only recently begun to incorporate big data into their analytic frames, and sociology in particular has been slow to adopt big data approaches. Of the several barriers that social scientists face, one is that big data originates from other disciplines, so has definitions of data and methods that are new. Sociologists are creating their own definitions for big data that are useful for their research purposes. Further, ethical questions that have long framed scientific research are being explored with this new type of data, as social scientists try to find ways to use data that has to potential to expose private identities, and that typically does not allow for the process of asking for informed consent.

KEYWORDS

i523, hid347, social sciences, big data, definitions, ethics, critical data studies

1 INTRODUCTION

The social sciences have only recently begun to incorporate big data into their analytic frames. While it may seem that sociology would be a prime fit for big data given the scope of their discipline–describing and theorizing all of society–there have been relatively few thorough explorations of how to apply big data to sociological analysis, particularly in the top sociology journals. Despite this, outlets have appeared that have taken a lead role in publishing the overlap in these two fields, and solid work has begun.

Three issues will be addressed. First, how rapidly have social scientists been incorporating big data into their research and publishing? Second, what are sites of overlap between sociology and big data, and how is big data defined in terms of what sociologists study. Third, what ethical questions have arisen for sociologists about the usage of big data, and how are these issues being addressed?

2 TRACKING SOCIAL SCIENCE USE OF BIG DATA

The use of big data to do social science analysis is relatively new. Using the Google Scholar index to track specific terms by year creates a picture of the rapidity with which social scientists seem to be exploring big data. Prior to and inclusive of 2005, Google Scholar had 7,560 records containing the phrase, *big data* (excluding patents and citations; retrieved 10/5/2017). Figure 1 shows the cumulative number of records in Google Scholar from 2005-2016 that contain the phrase *big data* plus either *sociology* or *social science.* There were only 559 in 2005, a mere 7.4% of the total *big data* references up to that point. The tipping point seems to be around 2012-2013 when these terms start to appear together. While after 2015 the number seems to level off, that may simply be an artifact of Google Scholar not yet picking up references since then. Regardless, there has clearly been a dramatic surge in the last five years. It is unlikely that all of these represent primary research by social scientists using big data, but it does represent a significant increase in the terms being found in the same articles, implying intersections of interest between the fields.

The recency of sociology's usage of big data is particularly striking when seen through the sociology-specific database, Proquest's *Sociological Abstracts* (Figure 1). There are significantly fewer results, since it only indexes peer-reviewed articles in sociology journals, compared to Google Scholar's results with the diversity of their indexed sources. From 2005-2016, there were 517 total references to *big data* in their 1,800 indexed serials (retrieved 10/7/2017), with almost all of those since 2013–there were only 2 references from 2005-2011. From what had been indexed as of October 2017, there were already 157 references for that year.

Academics publishing in the top sociology journals seem not to be using big data techniques with significant regularity. The ISI Web of Science tracks impact factors for peer-reviewed journals, and those values can be used to create a general (if not somewhat controversial) list of the top journals in any given field. Based on the impact factors for 2015, the top ten journals in sociology have a total of 92 usages of the term *big data*, according to a Google Scholar search (10/5/2017). This is a total search of any timeframe, and not all of these references represent primary research using big data, but simply refer to the usage of the terms. The top three journals each have 17 usages of the term, the most of these top ten, while Social Problems contains only 1 reference. Figure 2 shows these top ten journals along with a count of the usage of the term *big data* from the Google Scholar search.

In contrast, a relatively new journal began publishing in mid-2014 by Sage, Big Data and Society (BDS). It self-describes as publishing "interdisciplinary work principally in the social sciences, humanities and computing and their intersections with the arts and natural sciences about the implications of Big Data for societies." While primary research using big data in traditional sociology journals is relatively sparse, BDS publishes twice a year, containing primary research, and other relevant discussions, such as ethics and research methods. Because of its specificity, it is an important resource for the overlap of these fields.

3 DEFINING BIG DATA

Big data has been described as having velocity, variety, and volume, attributed to Laney in the early 2000s.[7] Kitchin includes five additional concepts: exhaustive in scope, fine-grained resolution, relational, extensional (ability to add fields), and scalable (the latter two concepts sometimes combined under the single concept of

flexibility).[8] In this analysis, Kitchin argues that big can be differentiated from small data specifically along these eight axes, and that while some small datasets may have characteristics of big data, such as strong relationality or wide variety, there is little overlap along the other axes. In their 2016 review, Kitchin and McArdle test 26 datasets that have previously been defined as big data for their *fit* based on these differentiating concepts.[9] They conclude that not only do not all datasets fit all of the criteria Kitchin described, but they also do not all fit the original descriptions of volume, velocity and variety. However, they do believe that all of the 26 datasets are characterized by velocity and exhaustivity, which they describe concisely as, "real time flow of data across a whole system" that produces a large dataset. The other descriptive concepts are still relevant, and may be pertinent to some big datasets, but not to others.

Others have taken different approaches to understanding what big data is. Dalton, using a political sociology lens, notes that the size of the dataset is clearly not the defining issue, since administrative data, such as the Census, can easily contain millions of subjects with thousands of variables, and this is not typically classified as *big data.*[4] Instead, he notes that while data like the Census is constructed specifically for the researchers' purposes, *big data* is often indirect, having been collected from *unconventional sources* and often involves merging several distinct datasets. In this case, finding similar types of data for cross-national studies can be a challenge, but typically involves finding government, administrative data on similar topics in many countries that have the possibility of being merged for comparative analysis into one dataset. Similarly, Brayne describes that sociological research shifts the definition away from specific characteristics of the data itself, like the size or speed of the data, but on the various institutional sources from which the data is gathered and then merged, thus emphasizing the process of collection rather than data features.[2]

Connelly, et al, and Japec, et al, discuss the *found* nature of big data in sociology, also highlighting the importance of administrative data, and the issues of merging data that was often constructed by different people, in different times, and for different purposes, but that may be similar enough for comparative analysis.[3] [7] Additionally, they note other types of data, particularly social media data, which can describe the real-time behavior of users. This type of data, unlike surveys, experiments, or other types of *researcher constructed-data* was not necessarily ever intended to become research data. But since sites like Twitter and Facebook collect data from their users and make it available, or when researchers can scrape data directly as it happens, it becomes information that can be re-purposed by social scientists. Continuing along that line, they describe two types of data–made and found. The former, which has been the typical source of social science research data, consisting of experimental and observational data, does not typically have the features associated with big data, but is designed to answer specific questions, and is highly systematic. Found data, on the other hand, whether administrative, or otherwise, is often very messy when it comes to the researcher, and is often not systematic, since it was not intentionally designed to answer a research question.

Yet another approach to understanding sociological types of big data is to look at the outcome of using it. In contrast to the previous approaches, which focused on qualities of the data or the way it was produced, Madsen, et al explore how big data functions to change society and the people it impacts.[11] They look at three aspects of big data using this frame: how it satifies daily life, how it can produce new modes of governance, and how it can be used to predict. Arguably, all of these have been occurring for a very long time, particularly since governments have been formally collecting data on its citizens. However, they argue that the large-scale collection of real-time data, and the integration of disparate sources of data has the potential to dramatically intensify these processes. Brayne highlights each of these concerns with her analysis of the data collection procedures of the LAPD.[2] She provides the example of ALPR, the automatic license plate reader system, by which cameras from various city sources, including those on police cars, automatically capture all license plate numbers it detects and logs them as geotagged information. First, mass numbers of citizens' locations, or at least their cars' locations, have been *datified*–where they were at a specific time that it was caught on camera. Second, this represents a novel form of surveillance by the state. In previous times, people were only logged into a policing system if they had been stopped because they were reasonably suspected of criminal activity. ALPR represents a wide net being cast to sweep up all people, now placed into policing databases, regardless of legal status of probable cause or consent. For Brayne, this represents a fundamentally new form of relationship between citizen and state, a new form of governance. Third, these and other types of data collection techniques have transformed traditional policing models at LAPD to predictive models, in which law enforcement is not simply responding to citizen calls, but generates incentives for the LAPD to invest resources into specific districts that algorithms have determined are more likely to experience criminal activity. Like the issue of wide-sweeping nets that put citizens into policing databases, predictive policing also represents a new modality of governance, since it directs public funds and interactions between the state and the public.

4 ETHICAL ISSUES OF BIG DATA IN THE SOCIAL SCIENCES

These issues lead to the questions that social scientists have raised about the ethics of using big data for research. After several classic ethical failures on the part of social and medical scientists, such as the Tuskegee syphilis study, Zimbardo's prison study with Stanford college students, and Milgram's electric shock experiment, the scientific community, along with government agencies, came together to create the Belmont Report (1978), outlining fundamental principles of ethical research, as well as providing specific ways to apply those principles.[6] Each of the principles on their own has proved resilient as a way to plan ethical research with humans: respect for persons, beneficence, and justice. However, these can be abstract and difficult to conceptualize their application, particularly in boundary or novel situations. Big data represents such a challenge.

As Brayne highlights, state usage of big data leads to legal questions about the relationship between the state and citizens. However, as researchers many of the same issues she raises are relevant. One of the specified applications of the Belmont Report is

the requirement for informed consent from research participants– researchers are expected to fully inform participants of risks and benefits of a study, the goal of the study, assure them they can leave the study at any time, and then get formal consent to include the individual in the study. However, mass collection of public or administrative data for research purposes almost never follows this guideline. To balance difficult cases, ethics committees often consider the risks to the participant when ideal protocols cannot be followed. But these can be difficult to assess, especially in new research territory, with new forms of data. McFarland, et al, note that privacy and potential harm to those caught up in mass data sweeps are of particular concern. [1]

Lazer and Radford discuss these same ethical concerns. [10] They note that while some of these issues are not new, for example, the common practice of using government administrative data, such as the Census, for research purposes, are largely considered resolved, since strict protocols are in place to prevent any individual from being connected to their specific information. In that case, while there is no informed consent by the subject for the researcher to use the data, there is little chance for harm to the individual, and little chance for privacy breaches. The repository of that data, the federal government, is responsible for data security and data de-identification in that case, not the researcher. Other large datasets, such as surveys implemented by the National Institutes of Health, or the General Social Survey, take similar extraordinary precautions to prevent data breaches that connect survey responses or medical histories to individuals.[10] On the other hand, the *found* nature of big data provides very different challenges. For example, for data sweeps that pick up Facebook or Twitter data, it may be very easy to connect individuals to information since much of that information is publicly available. Fiske and Hauser describe recommendations made by most Institutional Review Boards, which includes requiring researchers to specify a privacy-protection plan that has to be registered with the IRB to ensure subject's privacy faces minimal risk of exposure. Reuse of such data would continue to require identities of subjects to be protected. Using these types of publicly available data would be classified as *excused* research, meaning issues of informed consent would be considered waived, as long as the research subjects' identities were protected. [6]

Several sources point to the need to incorporate *critical data studies* into big data research, particularly research that incorporates products of human behavior, like social media or administrative data. Felt (2016) describes issues of social science research with social media. [5] They note the number of studies that use data from specific sources, specifically Twitter and Facebook, or some combination of those and other sources. While they note the specific privacy issues of connecting information from individuals across various platforms, they discuss the broader framework that social scientists should employ to think critically about how the data is being used by researchers. In contrast to previous eras, when data was often considered neutral, Felt argues that the use of social data cannot be considered neutral since it has the capacity to impact the subjects whose data is being used, for the research findings to retroactively come back to the subjects. Critical Social Theory from the early 1900s builds the foundation for this kind of interrogation of the use of the data, how the data was generated, and the types of research questions that were originally asked that led to this

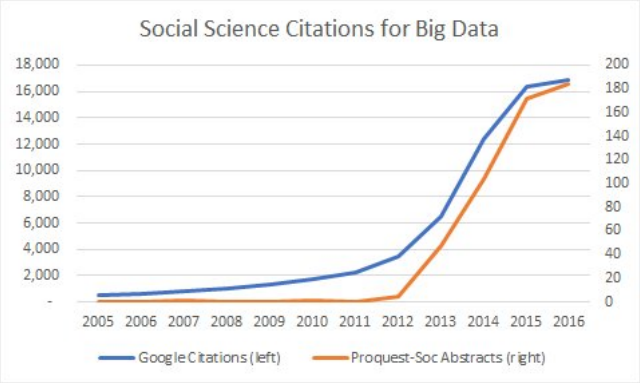

Figure 1: Citations over time for the phrase Big Data plus sociology or social science in Google Scholar (blue, 10/5/2017) and citations for the phrase Big Data in Proquests's, Sociological Abstracts (orange, 10/7/2017).

data being used. For example, they point to the argument that data is inherently political, and urge the researcher to ask whose interests are being served by generating and using this data, and to particularly be sensitive to how inequalities may be maintained or exacerbated by the data. Further, they ask the researcher to consider broader relationships between the data and society, highlighting the fact that data is never *raw*, but is always constructed, even if it is scraped from existing sources, and that the data represents specific power relationships within society. They urge an orientation that intentionally looks for socially progressive ways to think about and use data.

5 CONCLUSION

Sources and processes of big data are relatively new, and the social sciences are trying to catch up to the new resources they offer, and challenges they pose. Sociologists, in particular, seem to have only started publishing research using big data in the last five years when a rapid climb becomes evident in the database *Sociological Abstracts*. One task for social scientists is to define what constitutes big data. Research in physics, such as work to discover the Higgs-Boson, where hundreds of trillions of collisions had to be analyzed, far outpaces the size of data that social scientists would historically have analyzed, and likely continues to be exponentially larger than data they would analyze today. Sociologists describe administrative data as big data, particularly in the way that disparate datasets may be combined, with the complexity and relationality that entails. They also describe real-time data such as that collected from social media. Another early-stage set of questions with which researchers need to grapple are ethical concerns about the subjects they study, and the impact on society of the use of big data research. The privacy and safety of participants need to be protected, and issues raised by critical data studies urge researchers to ensure that research has a positive impact on those that might otherwise be disenfranchised from social resources.

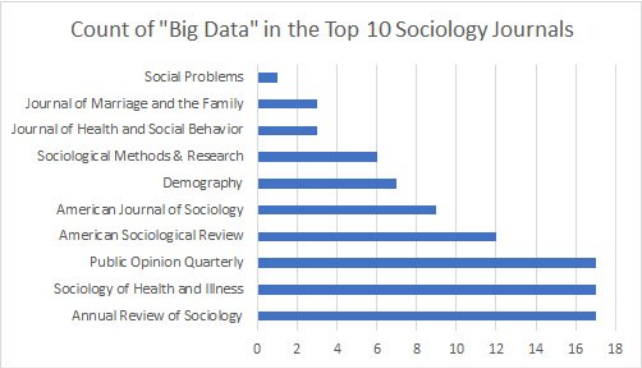

Figure 2: Count of the phrase Big Data in the top 10 sociology journals. Google Scholar, 10/5/2017

REFERENCES

[1] Daniel McFarland A., Kevin Lewis, and Amir Goldberg. 2016. Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *The American Sociologist* 47, 1 (01 Mar 2016), 12–35. https://doi.org/10.1007/s12108-015-9291-8

[2] Sarah Brayne. 2017. Big Data Surveillance: The Case of Policing. *American Sociological Review* 82, 5 (2017), 977–1008. https://doi.org/10.1177/0003122417725865 arXiv:https://doi.org/10.1177/0003122417725865

[3] Roxanne Connelly, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research* 59, Supplement C (2016), 1 – 12. https://doi.org/10.1016/j.ssresearch.2016.04.015 Special issue on Big Data in the Social Sciences.

[4] Russell J. Dalton. 2016. The Potential of Big Data for the Cross-National Study of Political Behavior. *International Journal of Sociology* 46, 1 (2016), 8–20. https://doi.org/10.1080/00207659.2016.1130410 arXiv:http://dx.doi.org/10.1080/00207659.2016.1130410

[5] Mylynn Felt. 2016. Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society* 3, 1 (2016), 2053951716645828. https://doi.org/10.1177/2053951716645828 arXiv:https://doi.org/10.1177/2053951716645828

[6] Susan Fiske and Robert Hauser. 2014. Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences* 111, 38 (2014), 13675–13676. https://doi.org/10.1073/pnas.1414626111 arXiv:http://www.pnas.org/content/111/38/13675.full.pdf

[7] Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy OfiNeil, and Abe Usher. 2015. Big Data in Survey ResearchAAPOR Task Force Report. *Public Opinion Quarterly* 79, 4 (2015), 839–880. https://doi.org/10.1093/poq/nfv039

[8] Rob Kitchin. 2014. Big data, new epistemologies and paradigm shifts. *Big Data and Society* 1, 1 (April 2014), 1–12. https://doi.org/10.1177/2053951714528481

[9] Rob Kitchin and Gavin McArcle. 2016. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society* 3, 1 (Jan. 2016), 1–10. https://doi.org/10.1177/2053951716631130

[10] David Lazer and Jason Radford. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43, 1 (2017), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457 arXiv:https://doi.org/10.1146/annurev-soc-060116-053457

[11] Anders Koed Madsen, Mikkel Flyverbom, Martin Hilbert, and Evelyn Ruppert. 2016. Big Data: Issues for an International Political Sociology of Data Practices 1. *International Political Sociology* 10, 3 (2016), 275–296. https://doi.org/10.1093/ips/olw010

Big Data and Deep Learning

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

ABSTRACT

Big Data is providing new opportunities for various industries in different sectors to enhance their performance by performing analysis on the huge amounts of data available. However, this is not as easy as it is said. Storing, transforming and performing analysis on such large amounts of data requires good storing and computational power. Many solutions have been proposed to handle big data and use it to the benefit of the company like Map Reduce, Spark and so on. Deep learning is one of the popular branches of machine learning which plays a key role when it comes to big data analytics.

KEYWORDS

i523, hid208, Deep Learning, Big Data, Deep Belief Network, Convolutional Neural Network

1 INTRODUCTION

Big data is the latest hot topic in the technical world and so is deep learning. Any data which consumes more than 1 Terra Bytes of memory is considered as big data. Social Media websites like Facebook generate more than 500 Terra Bytes No conventional data base cannot store or manage more than 1 Terra Bytes of data. Hence, new technologies like Hadoop, Spark and so on have emerged to store and process large amounts of data. Hadoop used HDFS for storing the data and Map Reduce for processing the data. Scripting languages like Pig, Hive and Spark can also be used to process the data but, Map Reduce is a better option for processing unstructured data.

Deep Learning is another hot topic which is being discussed almost everywhere. Deep learning is one of the branches of machine learning, which uses machine learning techniques to solve the problems of data analysis and prediction. It does not follow any pre-defined algorithms, rather learns from the data. The learning can be supervised or unsupervised. Deep learning is used along with the systems with high computational power to address the big data problems. Many companies like Facebook, Apple, Google, Samsung are using the deep learning techniques to manage the huge amounts of data that is being generated daily by their search engines and websites. Not only this, deep learnig is also used in speech recognition, image processing, weather forecasting and so on. Hence, the voice assistants like Siri, Google home, Alexa make use of deep learning as well. As the data keeps getting huge, deep learning comes into play to process the data.

2 DEEP LEARNING

Deep learning learns multiple levels of the deep architectures such as Deep Belief Networks(DBN), Convolutional Neural Networks(CNN) and so on. In this paper, a brief overview of DBN and CNN is given.

2.1 Deep Belief Networks

Any conventional neural network can only learn from the labelled data. But, most of the big data available is unlabelled. To take advantage of this massive amounts of unlabelled data, deep belief networks are used. They can not only learn from the labelled data, but also from the unlabelled data. They use both supervised and unsupervised learning techniques. It uses unsupervised techniques for pre- training and then to tune the data, it uses supervised techniques. Figure 1 shows the architecture of DBN.

To achieve this, DBNs use Restricted Boltzmann Machines(RBMs). RBM consists of input layer, hidden layer and an output layer. Nodes in each layer are connected to all the nodes in the adjacent layer(input to hidden) and nodes in same layer are not connected to each other. Hence, we can say that nodes in same layer are independent of each other. The nodes in hidden layer are connected to the nodes in output layer according to the output to be generated. The network is pre-trained layer by layer using unlabelled data and the generative weights of each RBM are found using Gibbs sampling[7]. The output of an RBM is fed as input to the RBM in the next layer. This process is repeated until all the RBMs in a network are pre-trained. The weights represent the input data. Then, the output layer is constructed according to the required outputs. Then, fine tuning is performed using labels and by back propagation. RBMs can be trained on unlabelled and large amounts of data. The sampling probabilities of hidden and visible layers of an RBM with bernoulli distribution are as follows:

$$p\left(h_j = 1 \mid \mathbf{v}; W\right) = \sigma\left(\sum_{i=1}^{I} w_{ij} v_i + a_j\right) \tag{1}$$

$$p\left(v_i = 1 \mid \mathbf{h}; W\right) = \sigma\left(\sum_{j=1}^{J} w_{ij} h_j + b_i\right) \tag{2}$$

The weights are updated using the following equation. The $(t+1)^{th}$ weight is updated as:

$$\Delta w_{ij}(t+1) = c\Delta w_{ij}(t) + \alpha\left(\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model}\right) \tag{3}$$

2.2 Convolutional Neural Networks(CNN)

CNN is another multi-layer neural network which is used for deep learning. CNN consists of multiple layers of convolution, activation and pooling. Figure2 depicts the architecture of a CNN

Convolution is a mathematical operation on two functions. It is defined by the following equation:

$$s(t) = x(t) * w(t) = \sum_{a=-\infty}^{\infty} x(a) w(t-a) \tag{4}$$

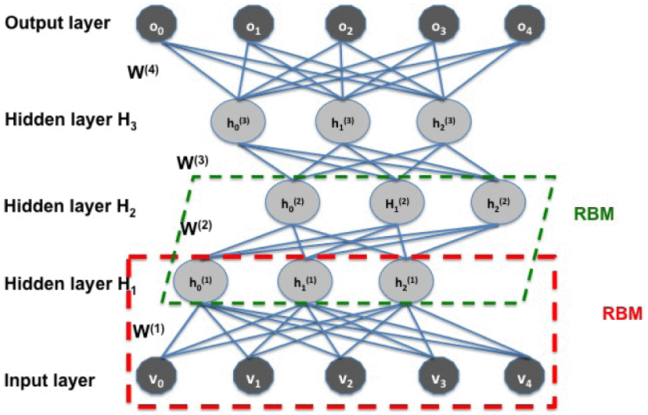

Figure 1: DBN

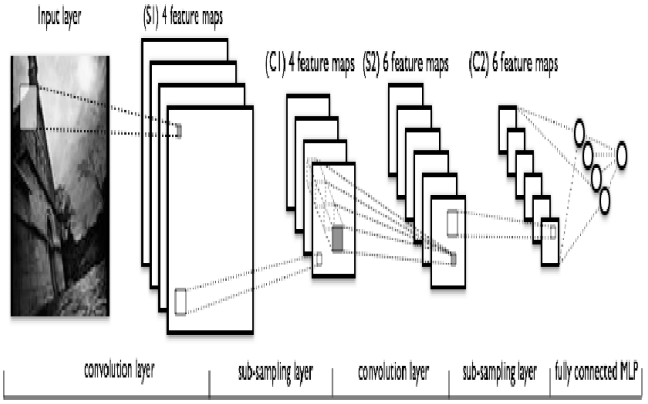

Figure 2: CNN

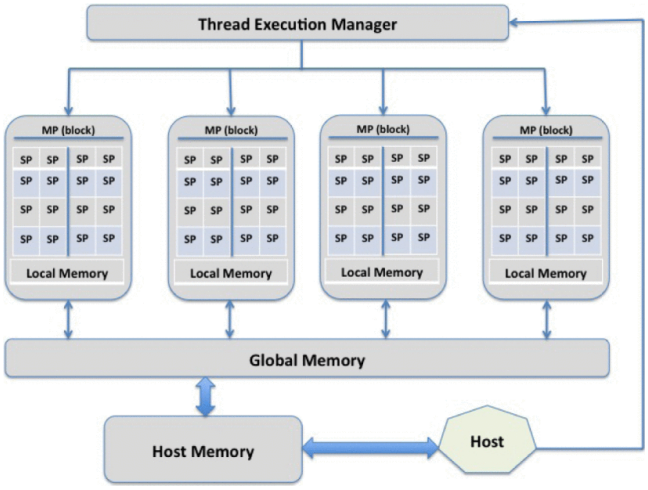

Figure 3: GPU

Convolution is used to extract the features from the input given to CNN, hence, it is called feature map stage. The output of convolution is called a feature map. The size of the feature map depends on three parameters:

(1) Depth
(2) Stride
(3) Zero-Padding

Depth is the number of filters used for convolution. Stride is the number of pixels used to slide a filter across the input during convolution and appending zeros for the sake of convolution is called zero-padding.

After the convolution operation, activation function is applied to the output, to introduce non-linearity. Any non-linear activation function can be used for this purpose.

After applying the activation function, the output is passed through pooling stage where different pooling operations such as average, sum, minimum, maximum, etc are applied. A window is defined and pooling is done in that window. The window slides over the input of the pooling stage by stride amount as discussed earlier.

After all these layers, a final layer is added using MLP for classifying the data or performing the regression, as per the required output. The steps followed in a CNN are, the filters and weights are randomly initialized first and then the data is passed through all the stages of CNN to get some output, then compare the actual and desired output and perform back propagation to update the weights accordingly.

3 DEEP LEARNING FOR BIG DATA

Deep Learning is very useful for prediction, especially when it comes to unlabelled data. It has proved to be efficient in many applications. But, when it comes to big data, deep learning algorithms are not very efficient, as training the nodes requires iterative computing of the weights which is very difficult when the data is huge. Hence, parallel algorithms must be used for training deep architectures, when dealing with big data.

In 2012, Deng et al.[6] proposed the concept of Deep Stacking Network(DSN) for parallel processing in deep architectures. Also there are other methods to address the big data problems, like improving the computation power, parallely computing the weights of hidden and visible layers, by distributing them across different machines and then integrating. It is nothing but a multi-node cluster in Hadoop. Each system is considered as a slave node computing the weights of a part of the hidden and visible neurons.

In addition to these techniques, systems with high computing power are used. One such example is GPU. Figure3 shows the architecture of a GPU:

The above GPU consists of four multi processors(MP), each MP consists of multiple streaming processors(SM) and each SM consists of multiple stream processors(SP). The stream processors share a common control logic and memory. The GPU also has a global memory. The architecture shown in the figure is a Single Instruction Multiple Thread architecture(SIMT). Such architecture is used when multiple computations are to be performed with less access time to the memory. The global memory in GPU is also a high-latency memory with high bandwidth. Here, the host represents the CPU. This architecture supports two levels of parallelism, namely memory level(MP) and thread level(SP). It facilitates multi-threading by running many hundreds of thousands of threads at a time.

3.1 Deep Belief Networks for Big Data

In deep learning architectures, millions of free parameters are considered to reduce the risk of over-fitting, in contrast to conventional architectures. For example Hinton and Salakhutdinov[8] have used 3.8 million parameters for images and Ranzato and Szummer[1] used three million parameters. But, the model proposed by Raina et al.[2] is far better than these models. Raina's model uses hundred million parameters for parallelizing the learning models which learn from unlabelled data, like DBNs.

Using GPU for parallelizing the DBN is not enough. Because, a considerable amount of time is wasted in transfering the data between the host and the global memories. Hence, to overcome this, a part of the training samples and the parameters are stored in the global memory itself while training. Also updating the parameters is done in GPU. In addition to memory and thread processing, data processing is also facilitated.

In DBNs, the weights are generated using Gibbs sampling using the same equations as for a non-parallel DBN, by generating sampling matrices P(x/h) and P(h/x) where the $(i, j)^{th}$ element is P(x_j/h_i) in P(x/h) and p(h_j/x_i) in P(h/x). Then, GPU is used to implement these two matrices. The weights are also updated in parallel using GPU.

3.2 CNN for Big Data

CNNs use GPUs for parallel processing to deal with big data. Both forward and backward propagation are used in training a CNN. Hence, both the propagations should be parallelized. To parallelize the forward propagation, each feature map in a CNN is assigned with some memory blocks, based on the size of the feature map and every thread in a block corresponds to a single neuron in a feature map. The CNN computations for each neuron in a map, such as convolution, applying activation function and pooling are performed in SP and the outputs are stored in the global memory.

In CNN, the weights are updated by back propagating the error. Back propagation can be parallelized by pushing or pulling the error signals. Although using GPUs facilitates parallel processing of data, it is only possible to process limited number of feature maps at any given time. For this purpose, Scherer et al[3] proposed an efficient method to use a circular buffer, which holds a small part of each feature map, loaded from global memory. Then, the threads parallely perform the convolution and the results are written back to the global memory. Krizhevsky et al.[4] proposed another yet faster method for processing big data using CNNs, by using two GPUs. Also, the speed of operation of CNN can be improved by using a ReLU or Rectified Linear Units activation function instead of any other activation functions.

4 DEEP LEARNING FOR HIGH VOLUMES, VARIETY AND VELOCITY OF DATA

The major concerns when dealing with big data are the volume of the data, variety of the data and velocity of the data. When dealing with large volumes of data, it is very difficult to train a deep learning algorithm using a single storage and CPU. Hence, distributed processing is preferred, which makes use of the multi-node cluster environment as in Hadoop. In such environment, the data and processing is distributed among different systems or nodes in the cluster for parallel processing and the outputs are again integrated at the master node.

Also, there are three types of data to be handled, structured, semi-structured and unstructured. Whatever the form the data might come in, it has to be stored and processed. There is not much difficulty in storing and processing structured data, but it is the semi-structured and un-structured data one faces a problem with. Also there is high velocity of data generated in many online site like the social media and so on, which needs to be accounted for in a timely manner. Deep learning can handle data of different varieties and with high velocity by using domain adaptation as discussed by Xue-Wen Chen and Xiaotong Lin.[5]

5 CONCLUSION

This paper discusses two of the available deep learning architectures and how they are used to to address the big data problems. Deep learning has proved to be useful whenever one encounters big data. Deep learning architectures can be used along with systems which have high computation power and by performing parallel processing.

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] 2008. *Semi-supervised learning of compact document representations with deep networks.*
[2] 2009. *Large-scale deep unsupervised learning using graphics processors.*
[3] 2010. *Evaluation of pooling operations in convolutional architectures for object recognition.*
[4] 2012. *ImageNet classification with deep convolutional neural networks.*
[5] Xue-Wen Chen and Xiaotong Lin. 2014. Big Data Deep Learning: Challenges and Perspectives. *IEEE* (2014).
[6] L. Deng and J. Platt D. Yu. 2012. Scalable stacking and learning for building deep architectures. *IEEE* (2012).
[7] G. Hinton and Y. Teh S. Osindero. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18 (2006), 7.
[8] G. Hinton and R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313 (2006), 5786.

Distributed Environment for Parallel Neural Networks

Ajinkya Khamkar
Indiana University
P.O. Box 1212
Bloomington, Indiana 47408
adkhamka@iu.edu

ABSTRACT

The past decade has seen the rise of Deep Neural Networks. A standard Deep Convolutional Neural Network has an upwards of Million parameters to train. Data required to train these networks typically ranges in Hundred's of Gigabytes, making it inefficient to train these networks on standalone machines. Graphical Processing Units decrease the computation time significantly but suffer from memory constraints. Existing Industrial architectures use a distributed computing paradigm capable of handling parallel computing tasks. We highlight approaches which use cheaper commodity systems integrated in a distributed fashion to handle training such Deep Neural Networks.

KEYWORDS

I523,HID211, Distributed Systems, Convolutional Neural Networks, Parallel Systems, Deep Neural Networks

1 INTRODUCTION

The past decade has seen the rise of Deep Neural Networks. Neural Networks have the ability to model complex non-linear functions by efficiently representing the input parameters as a system of linear equations with non-linear activation[3]. They have achieved unparalleled success in the fields of Computer Vision, Natural Language Processing and Artificial Intelligence. Section 2 discusses the number of parameters required to be trained for popular deep architectures. Large amounts of data is required to train these parameters. Section 3 discusses the size of the traditional data sets used to train these networks. Deep Neural Networks are inherently parallel in nature, with weights and gradient updates shared across layers within the network. Section 4 discusses various ways to introduce parallelism while training Deep Neural Networks. Section 5 discusses methodologies to update Model parameters when the data to train is distributed across multiple machines within the network. Section 6 introduces methodologies to train multiple layers of the same network in parallel in a distributed fashion.

2 POPULAR ARCHITECTURES

AlexNet[3], which achieved state of the art top 5 error of *19.80 %* for the Imagenet Large Scale Image Recognition Challenge in 2012 trained 60 Million parameters. In subsequent years, VGG-16 [4] a 16 layer deep convolutional neural network achieved achieved state of the art top 5 error of *8 %* in 2014 trained 138 Million parameters. ResNet [2], used a 152 layer deep architecture and trained 60 Million parameters to achieve top 5 error of *6.16 %*. DeepMind's Alpha Go agent ran on 48 CPU's and 8 GPU's.

3 DATA

Deep Neural Networks contain millions of parameters aligned across different layers within the network. They are designed to capture various features available in the input data. During training, the input data should cover maximum variance possible. This ensures that the network can generalize well on the test set and in real world scenarios. With respect to Convolutional Neural Network for image related tasks, this requires the dataset to contain equal representation of all classes that the network is trained to recognize or detect. Additionally the input data should contain images with varying object sizes and orientations to ensure the network remains translation and rotation invariant. Further, to ensure this network can generalize well in real world scenarios, the input data should contain common classes that maybe present in the environment. Thus we require millions of input images to successfully train a Deep Neural Network architecture. The Imagenet Large Scale Image Recognition dataset 2016, used to train popular deep learning classification algorithms has a total of 22,000 classes and has over 10 million hand-annotated images resulting in a dataset of size of 138 gigabytes. Youtube 8 Million video data set is one of the most popular datasets to train video classification algorithms. The total size of the dataset is 1.7 terabytes. Standalone devices and small device clusters are not capable of handling these datasets for training Neural Networks. These are stored in large memory clusters across several machines.

4 PARALLEL AND DISTRIBUTED ARCHITECTURES

4.1 Convolutional Neural Networks

Convolutional Neural Networks drive modern Computer Vision and Artificial Intelligence based research. The convolution operation involves sliding a filter of a predefined size over the input data and perform element-wise multiplication. They are capable of extracting higher level information from input data and project them to lower level embedding. The patterns identified in the lower level embedding can be used to perform various Machine Learning tasks such as classification, clustering, object recognition and source separation.

Parallelism of convolution operation. Every layer of a Convolutional Neural Network has a stacked input of filters. These filters are responsible for extracting higher level information from the input data. The filters operations are independently applied to the input data. This makes it possible to compute these operations in parallel to each other and collate their results[3]. Recent advanced software architectures such as tensorflow and theano are capable of achieving computation in parallel using multiple cores. Additionally Graphical Processing Units can be explicitly programmed

for parallel implementation of the convolution operator to achieve state of the art computational results.

4.2 Need For distributed approaches

Standard Convolutional Neural Networks have millions of parameters to train and optimize. Additionally the data required to train these systems ranges in Hundred's of Gigabytes. These computational constraints make it inefficient to train deeper networks on stand alone machines.

- Data Parallelism - When the data required to train neural networks exceed the systems storage capacity, it is required to distribute the data across multiple machines and introduce a data pipeline to feed input to the network[1].

- Model Parallelism - When the model being trained is too large to fit into the main memory. It is required to distribute different layers of the model across different machine and use distributed variants of Stochastic Gradient Descent to update each layer being processed on different machines[5].

5 DATA PARALLELISM

Data parallelism involves storing the input data required to train our Convolutional Neural Network Model across multiple machines. Each machine runs the same network model. Each model is then trained on an unordered random subset of the data. One of the biggest challenges faced in data parallelism is updation of model parameters. These are broadly classified into 2 categories.

- Synchronous update - In synchronous updates, gradients are computed using the loss generated by each model on a mini-batch of the independent input. Weights are updated using a single gradient generated by averaging the losses of each model.

- Asynchronous update - In asynchronous updates, each model runs independently. Global parameters shared by multiple models are held in a global parameter server. Each model then fetches the updated parameters from the server to process the mini-batch

5.1 Synchronous Updates

Zinkevich, Weimer, Smola & Li, 2010 [6] introduced a parallel variant of the traditional Stochastic Gradient Descent algorithm. They designed a simple yet efficient algorithm { see algorithm 1 } which averaged the gradients generated by the multiple machines within the network. This method is shown to converge and provide an optimal speedup. Algorithm 1 is applied iteratively either until convergence or until predetermined n epochs. An epoch corresponds to one pass of the model over the entire dataset.

5.2 Asynchronous Updates

Dean et. Al, 2012 [1] introduced an asynchronous variant of the traditional Stochastic Gradient. They proposed the use of a centralized communication server which holds parameters used by all models running in parallel. The communication server is distributed across several machines { see algorithm 2 }. Each model requests

Algorithm 1 Parallel SGD ($\{c^1, ..., c^m\}, T, n, w_o, k$)

for epoch $\in \{1....k\}$ *or* until convergence **do**
  for machine $\in \{1....k\}$ in parallel **do**
    compute feature maps
  $v_i = SGD(\{c^1, ..., c^k\}, T, n, w_o)$
$v = \frac{1}{k} \sum_{i=1}^{k} v_i$
Backpropagate v to update all model parameters in parallel

the centralized server for updated parameters before processing the mini-batch. Thus each model requests only those machines which holds parameters relevant to its partition. After computation of the gradient post processing the mini-batch the centralized server is updated with the new gradients. Subsequently the parameters are updated using the newly computed gradient. Asynchronous updates are more robust as compared to Synchronous updates. If a machine within the network fails, other machines are still up and computing their gradients. Algorithm 2 is applied iteratively either until convergence or until predetermined n epochs.

Algorithm 2 Downpour SGD (p, d)

for epoch $\in \{1....k\}$ *or* until convergence **do**
  for machine $\in \{1....k\}$ in parallel **do**
    query updated parameters from server
    $v_i = SGD(p, d)$
  Update centralized server with v_i
  $p = p - \nabla v_i$

6 MODEL PARALLELISM

Model parallelism involves training different layers of the Deep Neural Network in a distributed fashion across several machines in a network. In Model parallelism, different layers at the same level within the network are trained on the same input data. Model parallelism is required when the size of the network is too large to fit in main memory. Recent research in Deep Convolutional Networks is focused on the 'wider' paradigm instead of the traditional 'deeper' paradigm [5]. Wider Convolutional Networks can be viewed as a stack of smaller networks connected in parallel. Each of these smaller networks is designed and optimized to extract complex relationships in the input data at different depth levels. Wider Networks are computationally efficient than deeper networks. These smaller networks can be trained in parallel across multiple cores as these networks do not suffer from resource sharing. Each network in a layer gets its own copy of the output from the previous layer. A master layer is required to collate the results of the smaller networks to be passed to the next layer of the Network.

7 CONCLUSION

The number of parameters to train a neural network optimally have been increasing in the last few years. The data required to train these networks efficiently is continuously increasing. Standalone architectures are quickly being replaced with distributed architectures designed to handle training of these networks. Existing Industrial architectures can be tuned to train deep neural

networks. They are optimal for training such networks with little to no additional cost of setup and expertise. With the techniques presented above, deeper architectures can be trained efficiently and optimally to achieve state of the art results.

REFERENCES

[1] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. 2012. Large Scale Distributed Deep Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 1223–1231. http://dl.acm.org/citation.cfm?id=2999134.2999271

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[4] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. *CoRR* abs/1409.4842 (2014). arXiv:1409.4842 http://arxiv.org/abs/1409.4842

[6] 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

Big Data in Machine Learning

ZhiCheng Zhu
Indiana University Bloomington
936 S Clarizz Blvd
Bloomington, Indiana 47401
zhuzhic@iu.edu

ABSTRACT

With the development of IT technology, the world will enter the information age. People also called this "the era of third industrial revolution". Given the continuous development of the process of the third industrial revolution, all aspects of the traditional way of human society are changing. It can be said that every minute on the internet, have large amounts of new information be produced. With the increasing use of the internet and the increase of network bandwidth. People are making data every moment. It makes the information increasing quite quickly at a phenomenal rate. With so much data becoming available, getting data is not a problem for us anymore but find the right resource from the expanding information becoming a problem for most of the researchers. Because the data collected and stored at enormous speeds and human analysts may take weeks to discover useful information, traditional techniques are unfeasible for big data area. Therefore, we need to find new techniques to meet the challenge.

KEYWORDS

i523, hid229, Big data, Machine Learning, Technology

1 INTRODUCTION

The word "Machine Learning" first raised by Arthur Samuel, an American pioneer in the computer gaming and artificial intelligence areas [1]. In a broad sense, machine learning is a way to give the machine the ability to learn so that it can complete some task which not done directly by programming. It is a way of using Data to produce a model and then using the model to do prediction. Traditionally, if we want the computer to work, we will give it a series of computer instructions, and then the computer will follow this instruction step by step. The consequence always results by the code which you input before. And you can predict the result. But this way does not work in machine learning. Machine learning does not accept the instructions you entered at all, instead, it will accept the data you entered and applications of machine learning methods to these data sets and finally get produce a result. The core of big data is finding the value from the massive data sets, machine learning is a key technique that can effectively use the data value, for big data, the machine learning is indispensable. On the contrary, for machine learning, the more data will be more likely to improve the accuracy of the model. Therefore, the rise of machine learning is also inseparable from the help of big data. Big data and machine learning are mutually reinforcing and dependent.

2 CHALLENGES IN MACHINE LEARNING

Compared with the traditional machine learning, machine learning in big data can greatly expand the number of samples, the classification of many problems have a rich sample as support, this is the advantage of big data, but also caused many problems. Now, with the continuous optimization of hardware and programming algorithms, data collection and magnitude are no longer the major problems hindering the big data research. The relationship between different data sets, what kind of data is useful, which date is redundant and even cause interference to other data, how to reduce noise and make the model more accurate will be a challenge facing by machine learning when it mixes with big data. Big data has great potential value in all aspects of our society, it is not a simple task to obtain valuable information from big data. The core target of machine learning in big data is to excavate the data value which hidden in the data sets and find the information we need so as to maximize the value of data from the huge volume and structure of the data.

2.1 CHALLENGES IN PAST, NOW, AND FUTURE

Machine Learning has made extraordinary progress in the last 30 years. There have been significant advances in some area such as "Data Security, Finance, Marketing Personalization, Recommendations, and Smart Cars". But there are still a lot of obstacles for machine learning to go a step further. for example, lack of available data sets will be a problem whatever the past, the present, or the future. It has always been a problem. In the past, it is hard for a data scientist to find a tool to collect intensive data for researching. Nowadays, most of the data are collected by the big Internet company. the key problem of machine learning becomes to how people can get permission to access the data sets which owned by these big internet company. In the future, as the machine learning develop, the safety will become increasingly become the focus of attention. The Machine learning can be broken down into the following categories:

- "Supervised learning is a type of machine learning algorithm that uses a known data set (called the training data set) to make predictions. The training data set includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new data set. A test data set is often used to validate the model. Using larger training data sets often yield models with higher predictive power that can generalize well for new data sets." [3].

- "Unsupervised learning is a type of machine learning algorithm used to draw inferences from data sets consisting of input data without labeled responses" [4].
- Semi-Supervised Learning:A learning method combining supervised learning with unsupervised learning. Recognition is done using a large amount of unlabeled data and also using labeled data at the same time.

For example, one of the most common problem in Machine learning is Catastrophic forgetting, as we all know the machine learning need learning from the enormity of data and create a model. Catastrophic forgetting means the model will completely and abruptly forget previously learned information upon learning from some new data sets. If we want to achieve artificial general intelligence, then machine learning must be able to be used to perform multiple tasks. Even we can use representation learning and transfer learning to help us solve this problem to a certain extent but still has significant performance degradation. Another problem for machine learning is the safety. If we want to apply machine learning in peoplefis daily life. The security will be a question which the Data scientist unable to avoid. For example, in an image recognition test, "starting with an image of a panda, the attacker adds a small perturbation that has been calculated to make the image be recognized as a gibbon with high confidence"[6]. Another problem might raise because of the type of data sets. There are several different types of data in the Big Data area. The original unlabeled data and labeled data. the labeled will highly increase the efficiency of the process when the model starts to learning. Also, it will make the model more accurate when they do recognition. But the fact is even though the data increasing quite quickly at a phenomenal rate, with 2.5 quintillion bytes a day. but most of these data are unlabeled, which means these data are useless for supervised learning. and it also not suitable for deep learning which is the subset of machine learning.

3 APPLICATIONS IN MACHINE LEARNING

With the development expands and skills improved, machine learning becoming more popular and accepted by a lot of areas. Machine learning has been widely applied in data mining, computer vision, Natural Language Processing, biometrics, search engines, medical diagnosis and detection of credit card fraud, securities market analysis, DNA sequencing, speech and handwriting recognition, strategy games and robotics areas.

3.1 Machine learning in data mining

Data mining has been influenced by many disciplines, including database, machine learning, and statistics. To put it crudely, databases provide data management techniques, machine learning and statistics provide data analysis techniques.Many techniques provided by the statistical areas usually need further develop in the machine learning field, and then become effective machine learning algorithms before they can enter the field of data mining[5].Statistics affects data mining through machine learning, while machine learning and database are two major supporting technologies of data mining.

3.2 Machine learning in recommendation system

One of most common application in machine learning area is the recommendation system which running on the different Big internet company. Amazon may be taken as a typical example of the recommendation system. Based on a user's shopping record and a lengthy wish list, identifies which of the products the user is really interested in and willing to buy. Such a decision model can help the company to provide advice to customers and boost product consumption. for example, when you Log on to Facebook or GooglePlus, and they recommend the user who might be associated with you or you might know[2].

3.3 Machine learning in Marketing Personalization

According to the behavior pattern of the user during the free trial and the behavior in the past, which users might change to be a premium user, and which will not?. Such a decision model can help the company intervene in the program to convince users to pay sooner or better participate in product trials. For example, most of the video website and streaming media provider are willing to give user a free trial which can collect the user information and produce more attractive video or series to increase the user base[2].

4 CONCLUSION

Big data and machine learning are the two most popular fields in the Information Technology area. From the middle evil times' blocking of information to the explosion of data now, the amount of data in various fields and the scale of data sets have been increased at a phenomenal rate. The huge volume of data has brought huge potential opportunities and changes. With the proper use of the machine learning in big data can produce a lot of advantage. such as improve the efficiency. we can use the advantage of these data to help us make a better decision in different fields. one of a good example in scientific research is the data-driven research. In the scientific research, we can use the big data of the search engine to predict the ability widely used in the fields of medicine, astronomy and so on

ACKNOWLEDGMENTS

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

REFERENCES

[1] R. Kohavi and F. Provost. 1998. *Machine Learning*. Vol. 30. 271–274 pages.
[2] Bernard Marr. 2016. The Top 10 AI And Machine Learning Use Cases Everyone Should Know About. (2016). Retrieved October 09, 2017 from https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#7d4886ea94c9
[3] Mathworks. 2016. Introducing Machine Learning. (2016). Retrieved October 09, 2017 from https://www.mathworks.com/discovery/unsupervised-learning.html
[4] Mathworks. 2016. Introducing Machine Learning. (2016). Retrieved October 09, 2017 from https://www.mathworks.com/discovery/supervised-learning.html
[5] Margaret Rouse. 2017. data mining. (2017). Retrieved October 09, 2017 from http://searchsqlserver.techtarget.com/definition/data-mining
[6] Ophir Tanz and Cambron Carter. 2017. Why the future of deep learning depends on finding good data. (2017). Retrieved October 09, 2017 from https://techcrunch.com/2017/07/21/

why-the-future-of-deep-learning-depends-on-finding-good-data/

3

Big Data in Social Media

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

ABSTRACT

Big data refers to data that is generated from various sources including social media platforms. The impact of this data on marketing cannot be overlooked as many marketers have used big data analytics to create adverts that are tailored to the specific needs of users. This is possible because whenever a user comments or likes a product on twitter and Facebook, the same information is captured and used to analyze the customerfis behavior. Facebook uses Hadoop technology to help it dig deep into the customerfis behavior and draw important insights which can be used to suggest products for users based on their browsing patterns. This paper will evaluate big data and social media with specific focus on how big data, through social media, has been employed to improve marketing.

KEYWORDS

i423, hid109, Big data, Social media, Facebook, Harnessed, Platform, advertising, marketing

1 INTRODUCTION

Big data has been in existence for a number of years now. During this time several organizations have become aware of the fact that if they capture the streams of data flowing into their businesses, they can employ the power of big data analytics to get good value from it. Before this, basic analytics was being used in the early days as powerful analytical tools had not yet come into existence. The importance of big data in businesses cannot be overlooked as businesses have been able to draw valuable insights from big data analytics which has played a big part in helping them discover new opportunities. This approach has led to good business decisions, efficient operations, satisfied customers and increased profits [9]. In many cases the tremendous upsurge in the volumes of data can sometimes be overwhelming for businesses and individuals who may not know how to make sense of it all. However, with the shift in trend, an organization can adjust its strategy to help it compete favorably by using big data analytics. Subsequently, the most important concept to consider is the role that big data plays in social media marketing techniques. Social media sites, including Facebook, Twitter, and Instagram, are fundamental components of big data and form one of the most important sources of big data. The uninterrupted flow of content from various social media sites has enabled data collected from previous years to grow into big data.

2 SOCIAL MEDIA IS SIGNIFICANT FOR COMPANIES AND INDIVIVUALS

Although big data is said to come from several different sources, the largest proportion of it is said to originate from unstructured sources. As it can be imagined, social media makes up the largest source of unstructured content for big data. All the activities that users perform on social media such as views, retweets, comments, favorites, likes, etc. can be gathered and explored by interested individuals.

In the current digital world, social media plays a vital role in many companies. Having a presence on various social media platforms such as Instagram, Facebook, and Twitter is imperative since it enables individuals to interact with an organization on an ostensibly personal level and at the same time helps businesses across several domains get in touch with their customers. Currently, Facebook alone has over two billion users on their platform; this is roughly twenty-six percent of the world population [5]. It is therefore important to consider the fact that big data, from the social media platforms, can reach any people in different forms. Besides that, social media interactions have continued to play a big role and will continue to play a big role in business decisions. For example, some insurance companies have declined to offer life insurance policies to individuals solely based on their social media posts. If you frequently post, on any of these platforms, about how you are drinking or going to drink, insurance companies would be reluctant to offer you a life insurance policy as this is a risk to them.

It will not be long before organizations discover new and better strategies for making sense of big data. But, at the moment, the concept of big data is still new and rapidly evolving. Nevertheless, some businesses have found ways of interacting and using this data, which is just but the beginning, but still a good way to begin. To elaborate, a marketing company whose interest is promoting a new product could employ machine learning algorithms that enable it to gather data from individuals who meet certain attributes [5]. Consequently, by employing artificial intelligence technology, they will also be capable of drawing insights from millions of users and create campaigns. This will increase their levels of precision and focus, a technique usually referred to as targeted marketing, and present an excellent opportunity for finding the perfect audience and satisfy its preferences.

3 BIG DATA IN SOCIAL MEDIA ADVERTISING

Fundamentally, advertising revolves around communication since it is all about sensitizing consumers on products and services that an organization is selling. However, different consumers will always want to hear varied messages, which is a vital fact to consider when new clients are being recruited into the internet bandwagon due to the growing popularity of smart phones. Big data has the capacity to customize these messages, project what consumers would like to hear, and establish new perceptions on what customers like or

prefer [2]. The above steps are all revolutionary and are expected to have a significant impact on how marketers in various organizations advertise.

Furthermore, there are some occurrences which several people do not view as advertising but are still interactions between big data and marketing like product recommendation. An obvious example is Netflix [3]. Although the company does not have a concrete advertisement plan, it employees a lot of algorithms to recommend various movies and shows to its customers. The approach saves the organization a lot of money by reducing the rate of customer exit and ensures that the right shows are marketed to the right individuals. The companyfis strategy is to target consumers with shows specifically tailored for them. Apart from them, other firms such as Amazon, YouTube, etc. also do the same by using product recommendation to target their customers [3]. In order to stay up to date, the algorithms need constant flow of data to help it work more efficiently. With the growth of the internet, users leave huge volumes of data not only on social media platforms but also on other places they visit online in the form of a digital footprint. This provides advertisers with new avenues to tailor their messages to meet their customer demands.

The digital footprints left by online advertisers provides new insights to marketers on what a consumer really needs, and this sometimes may be more accurate than what the customer actually says on social media. However, marketers are worried about how to safeguard the privacy and security of their consumers and therefore companies that are careless in handling data collected from consumers usually ignite a backlash which greatly impact their business. Even though targeted advertising has been in existence for quite a while [3] the more the data that is collected by advertisers, the more personalized and effective marketing is expected to be. Organizations will strive not just to gather as much data as they can, but also to gather information which typically represents the individual consumerfis needs in order to enable them to market to their personalized tastes.

4 ANALYZING LINKS

Big data collected from social media can lead to the discovery of new information regarding each individual customer that can help in creating a customized appeal to that specific customer. However, with the new insights, marketers can enhance how advertising is approached as they create new strategies. The new growth in content marketing is usually perceived as a primary beneficiary of big data, although the concept of content marketing could be older than the internet itself.

Another essential point is that big data enables digital marketers to target users effectively with more personalized advertisements which they might prefer to see. Facebook and Google are among the biggest players in this domain of digital advertising. They have discovered excellent ways of creating and delivering more appealing advertisements in ways that do not intrude on the rights and preferences of the consumers [4]. Most of their advertisements feature services and goods that consumers would like most to enhance their lives and almost all of these advertisements are reliant on huge amounts of personal data that users usually provide from what they are up to, what they share and like things online.

Experts contend that it is possible to accurately make predictions on an array of individual attributes that are more sensitive merely through an analysis of an individualfis Facebook or Twitter likes [8]. For example, the likes on these social media websites are critical in predicting onefis religion, sexual orientation, emotional stability, life satisfaction, age, relationship status, and many other attributes. Companies like Facebook successfully linked political activity with user commitment when they created a sticker enabling most of their users to declare on their profiles that they had voted. The initiative was conducted during the 2010 midterm polls and was very effective as more people turned up to vote as compared to the 2006 midterm elections [6]. Individuals who saw the feature had high chances of voting and actively engaged in a conversation about the same after seeing their friends and peers participate in the activity. Later on, during the 2016 polls, Facebook escalated their role into the voting process by providing users with not only constant reminders but also with directions about their polling stations [7]. Apart from that, they also enabled users to easily get access to registration information, news, voting guides and other tools that would have made them more equipped to go through the election process.

5 USER RATING AND POP UP ADS

Depending on the user preferences and the content that they often access on social media, pop-up advertisements can be created to target users every time they are online. For example, an ad can be created on the Facebook Messenger app to open inside that particular app every time the user hits the CTA button. When clicked, such adverts would redirect the user to a page where they would be required to answer some question, claim a reward or send some feedback regarding a product or service. Before creating such ads, it is imperative to establish a custom audience of the individuals who would be targeted with that particular pop-up ads. For instance, individuals who have previously liked the companyfis products on their Facebook page or other social media sites can be included on the list of target audience to receive the ad [1]. Another strategy that can be employed is to rate users by tracking their cookies. In most cases, user activities are usually tracked across the internet using cookies whenever a user logs into one of the social media sites and is concurrently browsing other sites. Whenever this happens the other sites that the users is visiting can be easily tracked and the data used accordingly.

6 CONCLUSION

In summary, big data and social media have revolutionized advertising. Because of this, many organizations are currently harnessing the power of social media and big data analytics and creating marketing campaigns that target specific consumer needs. In the same way, different types of data generated from user interaction in various ways on social media platforms, such as comments on a status, likes, retweets, or shares, generate big data and enable the use of big data analytics. As a result, the data generated gives more insight about a userfis behavior and can be used to create marketing campaigns that specifically target users. Hence, this approach has

been proven to be effective and most efficient enabling most organizations to experience an increased customer growth and a boost in overall sales.

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] John Aycock. 2010. Springer Sciences & Business Media. *Spyware and Adware* 50 (2010), 71–109.
[2] Kyle Hensel & Michael H. Deis. 2010. Using Soical Media To Increase Advertising And Improve Marketing. *The Entrepreneurial Executive* 15 (2010), 87.
[3] Gary Eastwood. 2017. Big data, algorithms and the future of advertising. (2017). https://www.networkworld.com/article/3194585/big-data/big-data-algorithms-and-the-future-of-advertising.html
[4] W. Glynn Mangold & David J. Faulds. 2009. Social media: The new hybrid element of the promotion mix. *Business Horizons* 52 (2009), 357–365.
[5] David Geer. 2017. Will big data change how you use social media? (2017). https://thenextweb.com/contributors/2017/07/06/will-big-data-change-use-social-media/#.tnw_DPcEKg97
[6] Dara Lind. 2014. Facebookfis fiI Votedfi sticker was a secret experiment on its users. (2014). https://www.vox.com/2014/11/4/7154641/midterm-elections-2014-voted-facebook-friends-vote-polls
[7] Sarah Perez. 2016. Facebook gives its Election 2016 hub top billing by pinning it to your Favorites. (2016). https://www.qubole.com/blog/big-data-advertising-case-study/
[8] Nate Philip. 2014. The Impact of Big Data on the Digital Advertising Industry. (2014). https://www.qubole.com/blog/big-data-advertising-case-study/
[9] Dara Singh. 2016. Importance of Big Data for Social Media Analysts to understand customer impressions better. (2016). https://www.linkedin.com/pulse/importance-big-data-social-media-analysts-understand-customer-singh/

Big Data Application in Web Search and Text Mining

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

ABSTRACT

Because of the rapid development of social media, there are gigantic amount of data generated in every second on the web. And those data could be stored in any forms like text, videos, images or their combinations. The more complicated forms of data, the more space it will take up and will cost more time to read it. Although most of today's personal computers have a very high performance, it is extremely difficult to process and analyze useful text information from those huge amount of unstructured data by using traditional single computer methods without the help of big data tools or text mining techniques. Fortunately, the improvements in big data application are also increasing fast in order to support those difficult works on web search and text mining. This paper first studies the knowledge of web search technique and its data analytic steps, then introduces the link structure with a broad analysis of some web page structures (Hubs and PageRank), and at last, discusses their applications in this field of big data.

KEYWORDS

I523, HID209, Big Data, Social Media, Web Search, Text Mining, PageRank, Hubs

1 INTRODUCTION

In recent years, social media has become more and more popular as a new way of communication and knowledge transfer. People could use it to create, share, exchange information and create their own network. Social media usage has been boosted from 2005 to 2015. Users between 18 and 29 ages are the mainly part of social media users [7]. Today 90% of young adults are active on social media. This proportion was 12% in 2005 [1]. And since the development of mobile products, social media has also been offered a better platform for users to share data faster and more convenient. Thus, this proportion could be keep stable or still increase during the next few years.

Nowadays, a growing number of people prefer to express their opinion and feelings through tweeting, sharing images, commenting on social sites [7]. Since the amount of such data become extremely large, it is significant to extract and analyze useful information through them by using text analysis methods. Therefore, some applications which based on these information have been developed, such as recommendation system and search engine.

However, as the big data began to appear in the website, there are some problems that people must face for web search which include the longer search queries (key words) requirement, support the huge number of searches and multiple languages. And these problems cause the progress of web search and text mining technologies.

Web search is similar to information retrieval (IR) which is used to search for information on the World Wide Web [10]. The information may be a mix of web pages, images, and other types of files. Since web search is applying on web which has the huge amount of data, it has a much larger scale than many IR systems. Although web search is a complex technique, it has the capability to understand how to crawl internet to get and update information.

Text mining (also known as knowledge discovery in text database [4]) is semi-automatic process of discovering information, meaningful contents, topics, word, relations and patterns from a large amount of text data [7], which is also a branch of data mining. The text data could be extracted by web search at first.

2 WEB SEARCH TECHNIQUE

2.1 Key Fundamental Principles

Data, information, knowledge, and wisdom (DIKW) hierarchical model is the most basic model in the information management, information systems and knowledge management disciplines. Thus, it also used behind web search technique. It contains four main components as shown in its name: data, information, knowledge and wisdom. Since this paper only considers this model in web search area, these four components have the following conception.

- Data: raw web pages or "documents viewed as a bag of words".
- Information: result of query or "documents viewed as a collection of insights".
- Knowledge: result of processing query results by user.
- Wisdom: synthesis of many such actions by a set of users.

Figure 1 shows the hierarchical framework of DIKW model. It shows a pyramid structure with wisdom in the top level and data in the bottom level.

2.2 Search Engines

A web search engine is a software system for searching information on the Internet. The search results are generally presented in a line which are often referred to as search engine results pages. And some search engines also haeve the capability to mine data from databases or other open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain web crawling, indexing and searching processes in real-time [10]. Table 1 displays the development of search engines and some searching technologies in recent years.

2.3 Boolean and Vector Space Models

After discussed the basic principles and the application of web search, here introduce a model that used to define the search technique. Boolean model and vector model are both retrieval model

Year	Events
1990	First engine “Archie” appeared.
1994	Original Yahoo was human created catalog.
1995-2000	The classic information retrieval techniques adapted to HTML.
1998	Google founded with its link structure by using the PageRank algorithm.
2000-2005	Add context, spell check, suggestions, multiple sources.
2005-	Add optimization of complete results, topic analysis of documents, social search.

Table 1: Search engines development.

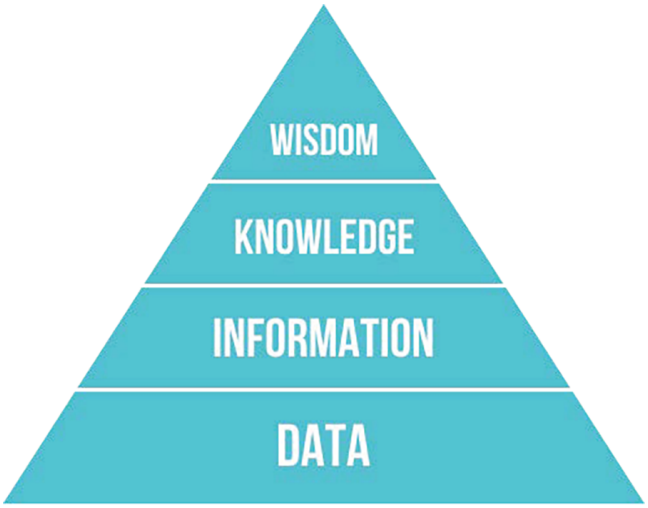

Figure 1: DIKW hierarchical model [8].

that can be a description of either the computational process or the human process of retrieval. For a retrieval model, it specifies the details of [6]:

- Document representation.
- Query representation.
- Retrieval function (how to find relevant results).
- Determines a notion of relevance.

In boolean model, keywords are considered to be either present or absent in a document and to provide equal evidence with respect to information needs. Queries are boolean expressions of keywords, which connected by AND ($\wedge$), OR ($\vee$), and NOT ($\neg$), including the use of brackets to indicate scope [6]. Thus, for the output of this model, the result document should be either relevant or not, and could not give partial matches or a ranking. Although this model is easy to understand and offers a clean formalism, it might become extremely complicated for most of web users in big data.

For vector space model, documents and queries are vectors in a high-dimensional space. Assume t distinct terms remain after preprocessing. Each term (i) in a document or query (j) is given a real valued weight w_{ij}. Therefore, both documents and queries are expressed as t-dimensional vectors [6]:

$$d_j = (w_{1j}, w_{2j}, \cdots, w_{tj})$$

There are some patterns to represent term weight. One is the Term Frequency, which assume that important terms have the higher frequency of occurrence in a document. The following equation define the vector space model.

$$tf(t,d) = \begin{cases} 0, & freq(d,t) = 0 \\ 1 + \log freq(d,t), & \text{otherwise} \end{cases}$$

While t refers to terms, and d refers to documents. This model straightforward to map everything to a vector and compare their angles. But it is hard to find a good set of basis vectors, a good weighting scheme for terms and a comparison function.

2.4 Web Crawling

It is not difficult to extract data from web with the help of algorithms. The input of the algorithm could be a list of URL's visited already and a list of new URL's to visit. Then executes the following steps in a loop:

(1) Fetch URL off list and check if done.
(2) If not done, go to web and continue collect.
(3) Hand document to document analyzer.
(4) Extract all URL's and add to list of new URL's to visit.

The result could be lots of detail of course. Then after fetching from the web, it should do the following steps:

(1) Convert document from HTML, PDF, Word, . . . to a text document.
(2) Tokenization: remove formatting, punctuation, capitals and convert to common form which makes document become a set of canonical tokens.
(3) Filtration: remove “stop words” (e.g. the, is, a, etc.).
(4) Stemming and Normalization: remove inflections and cope with non trivial synonyms.

Then the output are contents in bag of words and final terms are those used to define each dimension of vector space model.

3 WEB DATA (TEXT) MINING

3.1 Web Data Analysis Steps

For the big data which people search from web, it could be very difficult to extract or analyze useful information behind them. Thus, it is necessary to define the following steps to make those data structured or orderly so that people could easily applying other techniques like text mining to analyze them.

(1) Get the digital data from web.
(2) Preprocess data into searchable data like words or positions.
(3) Form Inverted Index in order to map words to documents.

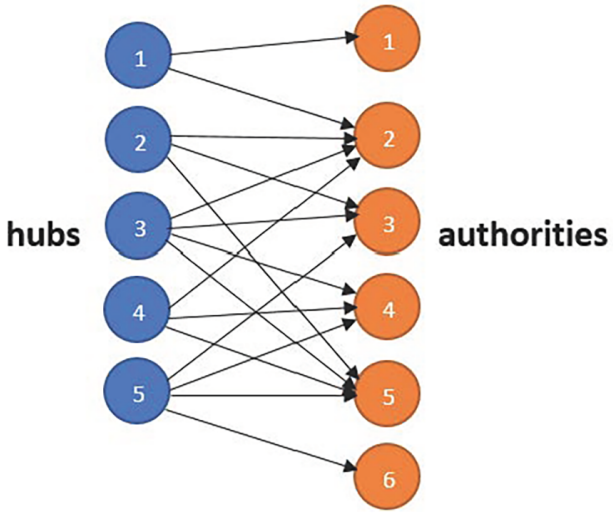

Figure 2: Hubs structure for web pages.

(4) Use algorithm like PageRank to rank relevance of documents.
(5) Apply some technologies (e.g. reverse engineering, preventing reverse engineering, etc.) for web advertising.
(6) Build the structure of the Internet and its people and pages.
(7) Clustering documents into topics.
(8) Might utilize Bayes to convert Mathematics of frequency into Mathematics of belief.

3.2 Link Structure Analysis

Since link structure has the significant impact to Search Engine Rankings, the PageRank flow and the number of pages that get indexed, it became one of the important factors of SEO (Search Engine Optimization) [2].

Link structure explores the connectivity patterns between web pages that contain the useful information and makes the huge of websitefis statistics meaningful. That is to say, mining these big data could help us understand what kind of things that users looking for, what are the hottest categories of a website and which pages are the most popular. Continuous optimization of link structure can eliminate duplicate content and promote popular pages in order to get more pageviews and higher rankings on Search Engine results [2].

An idea of the link structure for web pages is Hubs, which is known as Hubs and Authorities. The concept of this idea is simple: certain web pages served as large directories that were not actually authoritative in the information for users, but have links that led users direct to other authoritative pages [5]. Figure 2 shows the structure of Hubs.

As it shows in this figure, a good hub represented a page that pointed to many other pages and a good authority represented a page that was linked by many different hubs.

After defined the link structure of web pages, it comes to a link analysis algorithm named PageRank used by Google Search to rank websites in their search engine results. It is a way of measuring the importance of website pages. PageRank assigns a numerical weighting to each element of a hyperlinked set of documents with the purpose of "measuring" its relative importance within the set. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by $PR(E)$ [9]. The output of PageRank is a probability distribution that a page will be visited by a person who has the same probability to click each link on this page. This probability could be calculated iteratively with each page getting a contribution at each iteration equal to its page rank divided by the sum of links on page:

$$PR(\text{page } i) = \Sigma_{\text{page } j \text{ pointing at } i} \frac{PR(\text{page } j)}{\text{number of pages linked on page } j}$$

For example, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank [9]. PageRank could be used in ranking academic doctoral programs, recommendation systems and many other searching areas.

3.3 Clustering and Topic Models

After obtained results through a search query, it is important to classify them by groups for the further analysis. Clustering, also known as grouping document together, is the responses to a search query which give a group of documents. Suppose documents are the points in a space, the task of clustering is to identify regions. There are several ways to do this task:

- Clustering: Nearby regions of points.
- Support Vector Machine (SVM): Chop space up into parts.
- Gaussian Mixture Models (GMMs): A type of fuzzy clustering.
- K-Nearest Neighbors.

Alternatively, some "hidden meaning" can be determined with a topic model. It used to discover the abstract "topics" that occur in a collection of documents so that people could group documents by those topics. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body [3]. Assume each document is a set of topics and each topic is a bag of words, a topic model aims to find the best set of topics and best set of words in topics through a mathematical framework. That is to say, it allows people to examine a set of documents and discover what the topics might be and what each document's balance of topics is [3].

4 CONCLUSION

The aim of this paper is to demonstrate the core contents and the technique background of web search and text mining in big data area. Since the growth amount of data generate on web everyday cause the traditional computing methods and algorithms inefficiency, it is essential to make innovations in web search aspect. In the recent twenty years, search engines developed quickly and DIKW model, which was known as a popular model used in information system before, has applied in web for building its basic principles as well. As the vector space model appeared, the simple boolean model has been replaced in order to define the search query model more completely. And with the help of web crawling

algorithm, multiple types of text data extracted from website have become normalized before mining (analysis) the useful information.

Since webs page could seem as link structure, there must exists some patterns between linked pages. PageRank which found by Google is still widely applied in many different big data systems today, it has the ability to find the most relevant page for the content that the user searches for. After obtained pages of data information, we could utilize clustering to group documents together by topics.

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] Perrin A. 2015. Social Networking Usage: 2005-2015. (Octobe 2015).
[2] Bbriniotis. 2016. Link Structure: Analyzing the most important methods. (October 2016). http://www.webseoanalytics.com/blog/link-structure-analyzing-the-most-important-methods/
[3] David Blei. 2012. Probabilistic Topic Models. *Commun. ACM* (2012).
[4] Emir and Almir. 2016. Application of Big Data and Text Mining Methods and Technologies in Modern Business Analyzing Social Networks Data about Traffic Tracking. *IEEE* (October 2016).
[5] Christopher D. Manning and Prabhakar Raghavan. 2008. Introduction to Information Retrieval. *Cambridge University Press* (2008).
[6] R.Mooney, J. Ghosh, and D. Lee. 2017. Boolean and Vector Space Retrieval Models. (2017). http://www.cs.ucsb.edu/~tyang/class/293S17/slides/Topic2IRModels.pdf
[7] Mehmet U. and Secren G. 2016. Text Mining Analysis in Turkish Language Using Big Data Tools. *IEEE Computer Society* (2016).
[8] Wikipedia. 2017. DIKW pyramid. (September 2017). https://en.wikipedia.org/wiki/DIKW_pyramid#cite_note-Rowley-1
[9] Wikipedia. 2017. PageRank. (September 2017). https://en.wikipedia.org/wiki/PageRank
[10] Wikipedia. 2017. Web search engine. (October 2017). https://en.wikipedia.org/wiki/Web_search_engine

Big Data and Speech Recognition

Yuchen Liu
I523-HID:213
1750 N Range Rd, Apt D302
Bloomington, IN 47408
liu477@iu.edu

ABSTRACT

Nowadays, Speech Recognition is becoming more and more important. Many technology companies are trying to use Big Data to develop more efficient and accurate algorithm for Speech Recognition. Nowadays, Deep learning can be described as the foundation of Speech Recognition. Deep learning algorithms such as RNN and CNN often need to supported by large amount of data – Big data. Before Big Data and deep learning, the word error rate was 24 percent. Recently, IBM published a paper where the word error rate was below 5.5 percent. In August, Microsoft speech recognition system has reached a 5.1 percent error rate.

KEYWORDS

i523,HID213,Big Data, Speech Recognition

1 INTRODUCTION

Speech recognition, also known as automatic speech recognition, is a sub-field of computational linguistics. It can help computer to translate speech that we are spoken to text[13].

Speech recognition is a branch of pattern recognition, and it belongs to the field of signal processing science. It also has a very close relationship with phonetics, linguistics, mathematical statistics and neuro-biology. The purpose of speech recognition is to let the machine "understand" the language of human use. The word "understand" here including two meanings: one is verbatim to understand non-translated text into written language; the other is if the speech contained the request or ask, the computer will make the right response.

2 PROBLEMS OF SPEECH RECOGNITION

The definition of Speech Recognition has been raised for many years. Before 1980s, the most serious problem of Speech Recognition is the limited choice of algorithms. In 1952, the United States ATT Bell Labs developed the first electronic computer-based voice recognition system Audrey[12], which can identify 10 English figures, the accuracy rate of 98 percent .In the 1960s, the two major areas of speech recognition is linear Predictive coding and dynamic time specification techniques.

In the late 1960s, the Hidden Markov Model was proposed by Leonard E. Baum. HMM is a major breakthrough in the history of speech recognition. The error rate of speech recognition is greatly reduced[11]. HMM can be used in many areas. For example, in our daily life, we always want to predict the weather according to the current weather situation. One way is to assume that each state of the model depends only on the previous state. This assumption is called the Markov hypothesis, and this assumption can greatly simplify the problem. Obviously, this assumption is also a very bad assumption, which resulting in a lot of important information are lost. When it comes to the weather, the Markov hypothesis is described as assuming that if we know the weather information for some days before, then we can predict the weather today. Of course, this example is somewhat unrealistic. However, such a simplified system can be beneficial to our analysis, so we usually accept such assumptions, because we know that such a system allows us to get some useful information, although not very accurate.

In 1980s, artificial neural networks have been introduced into speech recognition[8]. Neural Network have many advantages. First, the neural network is non-linear. The neural networks that are interconnected by non-linear neurons are non-linear. The neural networks that are interconnected by non-linear neurons are non-linear in nature. Second, neural network is contextualinformational. The specific structure of the neural network and the state of excitation represent knowledge. Each neuron in the network is potentially affected by the global activity of all other neurons in the network. Therefore, the neural network will naturally be able to handle contextual information. However, because of the lack of high quality speech data and the lack of computational power, Neural Network still have a high error rate on Speech Recognition.

3 WHY BIG DATA IS IMPORTANT

Andrew Ng has predicted that as speech recognition goes from 95 percent accurate to 99 percent accurate, it will become a primary way that we interact with computers[4].In recent years, the idea of Big Data has been brought out. As the amount of data and computational power both increase, neural network is widely used in speech recognition tasks. Big Data becomes the answer of the problem of Speech Recognition.

There are more than 7000 different languages in this world and different people who speak the same language have different accent. Therefore, a large amount of data is required in order to make the Speech Recognition result accurate. A recent Google research paper shows that "Large language models have been proven quite beneficial for a variety of automatic speech recognition tasks"[3]. In the paper, the researchers found that data sets and larger language models will bring fewer errors predicting the next word based on the words that precede it. they also found that increasing the model size by two orders of magnitude will reduces the world error rate by 10 percent relative." The world error is 24 percent for Speech Recognition.

In March 2017, IBM announced that they are reaching a new record of Speech Recognition of 5.5 percent error rate. In order to get the goal, IBM combined LSTM (Long Short Term Memory) and WaveNet language models with three strong acoustic models[9]. The first acoustic models is a six-layer bi-direction LSTM model with multiple feature inputs. The second acoustic models is trained

with speaker-adversarial multi-task learning. For the third model, it not only learns from positive examples but also learn more the negative examples. So it gets smarter and smarter. It also performs better when similar speech patterns are appeared.

In August 2017, Microsoft then announced that they improve the Speech Recognition accuracy to 5.1 percent error rate. Basically, they improved the recognizer's language model by using the entire history of a dialog session to predict what is likely to come next, effectively allowing the model to adapt to the topic and local context of a conversation[10]. The data amount that they are using is huge and the improvement of the amount of data that they use result in a better result.

4 CURRENT ALGORITHM AND BIG DATA

Xuedong Huang, who leads Microsoft's Speech and Language Group, said that "People are speak with oxygen. Big data is just like the oxygen to speech recognition, there must be large data in order to make the algorithm accurate." He also said that for Speech Recognition, there are two things that are most important. One is data and the other is algorithm[6] .

A variety of neural network learning methods are in fact similar, basically through the gradient descent method to find the best parameters of the model. Then find the optimal model by deep learning. Nowadays, there are two different kind of neural network that is widely used in Speech Recognition field. One is RNN(Recurrent Nerural Network) and the other is CNN(Convolutional Neural Network).

The most important idea of RNNs is to make use of sequential information. In a traditional neural network we always assume that all inputs and outputs are independent of each other. But for many situations it is not true. For example, If there is a sentence said "Tom broke the glass, Mr.Peter critized ( )." In this situation, we should know that we need to put the name "Tom" in the blank. If you want to predict the next word in a sentence you'd better know which words came before it. RNNs are called recurrent because we can bring the information form the previous sentence to the next sentence, which the output being depended on the previous computations. In order to train a muti-layer RNN model, a large amount of data is required. Only if we let the model seen different combinations of text and different ways to talk, the accuracy of the Speech Recognition can be acceptable. Both Microsoft and IBM use a specific RNN architecture in their research, The model is called LSTM (Long short-term memory). The concept LSTM was invented by Hochreiter & Schmidhuber in 1997. It was invented to solve the Long-Term Dependencies of RNN[2].

CNN (Convolutional Neural Network) is very similar to DNN. They both build up by neurons and have a cost function. However, the input that CNN take is different from the Ordinary Neural Network. CNN architectures make an assumption that the inputs are images, which allows us to encode certain properties into the architecture. When a computer sees an image, it will see an array of pixel values. Depending on the resolution and size of the image, it will see a 32 x 32 x 3 array of numbers.For Speech Recognition, we can put a plot of the audio data as the input of the CNN model. The experimental results from Microsoft show that CNNs reduce the error rate by 6 percent-10 percent compared with DNNs on the TIMIT phone recognition and the voice search large vocabulary speech recognition tasks[1].

5 CURRENT PRODUCT AND BIG DATA

Speech recognition is becoming more and more important in our daily life. It is almost build into all our electronic devices. For example: phones, smart watches, computers and game consoles. It is even automating our home. People are becoming more and more familiar with talking to electronic devices. Many IT Giant is working on different algorithms to make the recognition result better by using Big Data. They also producing many interesting product: Apple Siri, Amazon Echo, Microsoft Cotana etc.

5.1 Apple and Siri

Although Apple is the most profitable technology companies, in the field of Big Data, it still catching up. However, they have increased the pace of entering Big Data. Recently, Big Data expert Bernard Marr analyzed how Apple used Big Data[7].

In the mobile market, Apple is a powerful presence. They have been actively encouraging developers to develop applications based on user data monitoring and sharing. An obvious example is that they recently announced a partnership with IBM to promote healthy mobile application development. They also developed a number of air travel, banking and insurance applications with IBM. Apple Watch's launch accelerated the process. Many commentators believe that smart watches may be the ultimate device for wearing devices. With more sensors, it can collect more data for more extensively analysis.

The most typical Big Data Application for Apple is Siri. In terms of speech recognition, the most common personal voice assistant in the United States has a really good performance. At 95 percent accuracy, Siri surpassed all of its Silicon Valley Silicon Valley giants. The most impressive part about siri is that it can perform really good in so many different languages. Even on some dialects. That is because of the use of Big Data. Only if the amount of data is large enough for the training process, apple can get this great result.

5.2 Google and Google Now

Google is the founder of the Big Data age. Its Big Data technology architecture has always been the study and research by other Internet companies. Google provides Big Data analysis intelligent applications such as customer emotional analysis, fraud analysis, product recommendation and Speech Recognition. Those Big Data applications has brought Google 23 million in revenue each day.

According to Mary Meeker's annual Internet Trends Report[5], Google's speech recognition model Google Now has achieved a 95 percent word accuracy rate on May 2017 for English, which is really similar of what Apple have. This current rate is also known as the threshold for human accuracy. This accomplishment is based on the huge data that Google could get every data. As the world mostly used search engine, Google gains more than 20PB of data everyday. This huge amount of data helps the model performance better.

6 CONCLUSIONS

In conclusion, Big Data is an indispensable part of Speech Recognition. Without Big Data, the existing powerful algorithms such as CNN and RNN will not work. These algorithms has gained huge successes in a broad area of applications now. Such as speech recognition, face identification, and computer vision. In today's world, the size of data that we can use is huge. Big Data also means big opportunities. Also, the huge amount of data also bring huge challenges to harnessing data and information. As the volume of data keeps getting bigger, using the proper algorithm plays a key role to increase the accuracy of Speech Recognition and other data analytic solutions.

REFERENCES

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (October 2014), 1533–1545. https://www.microsoft.com/en-us/research/publication/convolutional-neural-networks-for-speech-recognition-2/

[2] Rowel Atienza. 2017. LSTM by Example using Tensorflow. (18 March 2017). Retrieved October 2, 2017 from https://medium.com/towards-data-science/lstm-by-example-using-tensorflow-feb0c1968537

[3] Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. *Large Scale Language Modeling in Automatic Speech Recognition.* Technical Report. Google.

[4] Adam Geitgey. 2016. Machine Learning is Fun Part 6: How to do Speech Recognition with Deep Learning. (24 December 2016). Retrieved October 2, 2017 from https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a

[5] APRIL GLASER. 2017. Googlefis ability to understand language is nearly equivalent to humans. (31 May 2017). Retrieved October 2, 2017 from https://www.recode.net/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

[6] Microsoft Asia Research Institute. 2017. Huang Xuandong: How does Microsoft use artificial intelligence to do voice recognition? (24 April 2017). Retrieved October 2, 2017 from http://www.msra.cn/zh-cn/news/features/xuedong-huang-talk-20170424

[7] Bernard Marr. 2015. How Apple Uses Big Data To Drive Success. (22 May 2015). Retrieved October 2, 2017 from http://www.datasciencecentral.com/profiles/blogs/how-apple-uses-big-data-to-drive-success

[8] Margi Murphy. 2015. Everything you need to know about deep learning and neural networks. (19 August 2015). Retrieved October 2, 2017 from https://www.techworld.com/data/why-does-google-need-deep-neural-network-deep-learning-3623340/

[9] George Saon. 2017. Reaching new records in speech recognition. (07 March 2017). Retrieved October 2, 2017 from https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

[10] Catherine Shu. 2017. Microsoftfis speech recognition system hits a new accuracy milestone. (20 August 2017). Retrieved October 2, 2017 from https://techcrunch.com/2017/08/20/microsofts-speech-recognition-system-hits-a-new-accuracy-milestone/

[11] Asta Speaks. 2011. History and Theoretical Basics of Hidden Markov Models. (19 Apirl 2011). Retrieved October 2, 2017 from https://www.intechopen.com/books/authors/hidden-markov-models-theory-and-applications/history-and-theoretical-basics-of-hidden-markov-models

[12] Asta Speaks. 2014. Audrey: The First Speech Recognition System. (14 October 2014). Retrieved October 2, 2017 from https://astaspeaks.wordpress.com/2014/10/13/audrey-the-first-speech-recognition-system/

[13] Wikipedia. 2004. Speech Recognition. (March 2004). Retrieved October 2, 2017 from https://en.wikipedia.org/wiki/Speech_recognition

Using Big Data for Fact Checking

Karthik Vegi
Indiana University Bloomington
2619 East 2nd Street, Apt 11
Bloomington, IN 47401, USA
kvegi@iu.com

ABSTRACT

In this data age, the sheer volume of data makes it impossible to know what is truth and what is not. Politicians are often misconstruing facts to improve their candidacy. Scientists and advertisers are making false claims to gain business advantage. The more the false claims penetrate into the internet, especially social media, the more chances are that it is believed to be true. We show how Big Data techniques can be used to spot fake news, false claims made by politicians, advertisers, and scientists.

KEYWORDS

i523, hid231, big data, veracity, fact check, data accuracy

1 INTRODUCTION

Big Data is playing a crucial role in building a smarter planet. Each and every action that we take leaves a digital footprint. Big Data is lending a great helping hand to crunch this data and make smarter decisions. "*Big Data* is at the heart of the smart revolution, completely transforming the way we live, conduct science, run cities, and operate business" [6].

Analyzing data in this digital era where data can come from multiple sources involves reading data from different systems in different formats with different contextual meanings. The data extracted from multiple systems can often contradict each other. It could be biased towards a business or a particular entity. Multiple sources also mean conflicting and outdated information which makes it highly inaccurate [1].

Validation of facts became a major issue with the recent U.S. election of 2016 where the candidates from both the democratic and republic parties used a lot of factual statements in the debates to put their candidacy and party in a better position. These factual statements if not validated might give a false edge to the party thus having an effect on the entire nation. While some people take these statements with a pinch of salt, a large set of population often believes it to be true and end up voting for party purely based on the claims made by the respective candidates [2]. With so many data sources like social media, print media, and the internet, it is not easy to validate and spot fake news. We need to take the help of the technological advances in *Big Data* and *Artificial Intelligence* to handle this problem.

Data inconsistencies with respect to the sources, interpreting the data out of the context, obsolete data and data that is highly modified from the original are all data veracity problems [1]. Based on the discussion so far, fack-checking can be clearly identified as a data veracity problem which can be attributed to the fact that fake news and facts come from different sources, in different formats, often have missing factual details, and have inconsistencies [1].

2 FACT CHECKING AS A BIG DATA PROBLEM

Often veracity is not just about data quality, it is about data understandability. Fake news is understandable and we can make great sense out of it by careful analysis [2]. The data veracity problem in *Big Data* meets the fake news problem at the juncture called *Misinformation Dynamics* where the emphasis is not just on the inaccuracy because of an accident, but on the data quality as a whole [2]. Fake news is often intentional and moreover, it is not static but dynamic [2].

One straightforward way to understand and account for the reliability of the sources is to formulate a voting algorithm that labels the source system in which the data item resides thus evaluating the accuracy of the data [3]. The problem with this approach is that it is too simple and also it doesn't take into account the other factors such as the data lineage of each source [3]. This means that if multiple sources of the data are all derived from another source which is inaccurate, we are wrongly labeling the data source [3].

The social networking giants like Facebook and Twitter faced this problem and a lot of fingers were pointed at them for acting as a medium for spreading fake news. Facebook took the initiative to tackle the problem head-on by implementing an option where the users can flag the story as either true or false [2]. The more false votes s story garners, the less likely it appears on the news feed along with a warning message to the users mentioning that a lot of users have reported the story as false [2]. The problem with this approach is that we are giving people a chance to alter the truth, also making everyone believe that anything that is not flagged is true which might not always be the case [2].

In order to solve this problem in a more efficient way, we could combine *Big Data* and *Artificial Intelligence* techniques that eradicate the possible human-generated errors [2]. *Google* came up with a new method of scoring web pages based on the accuracy of the facts in which the algorithm assigns documents a trust score taking the context into account, which feeds the overall scoring to determine the search rank without solely relying on the links [2].

While the social networking giants have a huge role to play in identifying fake news, each individual should take personal responsibility to check the validity of the data using online tools at their disposal rather than believing it blindly.

3 BIG DATA TECHNIQUES FOR FACT CHECKING

3.1 Recommendation Based Approaches

Recommendation based approaches take the help of the community to determine the accuracy and quality of the sources [1]. The reputation of the sources increases as more people agree that the source

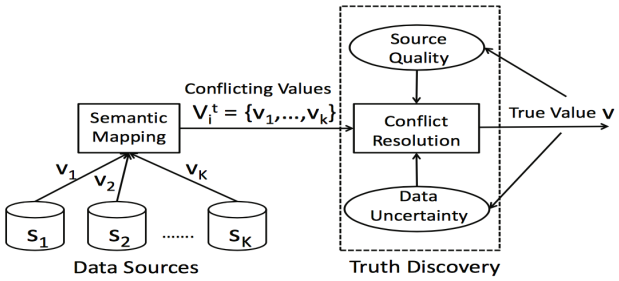

Figure 1: Truth Discovery In Data Streams [7]

is reliable [1]. These methods clearly have their shortcomings as people can be influenced by third-party agencies to improve the trustworthiness of certain sources [1].

3.2 Content Based Approaches

Content based approaches work by coming up with a score to compute the trust of a source and validates the belief of the claims it is making to generate a belief score [1]. The trustworthiness of the source now becomes a function of the trust score and the belief score [1]. This is not a one time process but the function runs over and over to continuously update the source quality [1]. A few probabilistic methods are added to this function to improve the accuracy of the score which extends the algorithm beyond just trust and belief [1].

In one such application, the truth discovery problem is transformed into a probabilistic inference model [7]. An iterative algorithm is proposed which computes the posterior distribution of all the values of the sources and finds the one with the maximum probability [7]. The model derives all the possible values reported by the sources and the conflicting values in the data streams and then calculates a score [7].

Figure 1 illustrates the content based approach for truth discovery in data streams. As there can be heterogeneous sources, first a semantic mapping is employed for the values provided by various sources such that the values for truth discovery are consistent [7]. Taking an example, the weather conditions that imply the same meaning such as *rainy* and *wet* are considered to be the same in truth discovery [7]. The same way, *partly sunny* and *cloudy* are considered as *clear* weather condition [7].

On each iteration, the fact-checking system collects all the links with conflicting information from different source systems for analysis [7]. The system then tries to remove the conflicts to arrive at the real true value taking the help of quality and authenticity of the source [7]. Finally, the system updates the computed value for each source and moves on to the next iteration [7]. The iterative process continues until none of the values change further [7].

3.3 Evidence Based Approaches

Evidence based approaches augment the content based approaches by relying on evidence, context and prior knowledge about the data sources [1]. Data provenance information may be used in truth discovery computation, as well as external information about the context, the sources, the data or user network [1]. This involves checking the dynamics of information in the network and recomputing the truth discovery accordingly [1]. Not every industry has a separate budget for research which makes evidence based approaches a viable option only for big organizations.

4 REAL-TIME FACT CHECKING

In this digital age, fact checking makes more sense when it is done in real time. Politicians and media houses use inaccurate facts and make claims and get away with them in real time, but the new fact-checking tools can often expose claims that are invalid and inaccurate [4]. The number of active fact-checking websites has been growing immensely, from about 44 in 2014 to about 114 currently [4].

The delay window between the time when a claim is made and the time when the claim is checked for truth has to be as less as possible as fact checking often takes longer time than traditional journalism [4]. This gives enough time for the politicians and other people to make a claim and get away with it [4].

4.1 ClaimBuster

ClaimBuster is a system that is built for real-time fact checking using the techniques of natural language processing and machine learning combined with database queries [5]. Although the complete system is still in works, some components of the system are already in use [5].

Figure 2 shows the system architecture of *ClaimBuster*. The *claim monitor* integrates the various data sources and feeds them into the system [5]. The *claim spotter* picks the claims that could potentially be checked for accuracy and reads in the relevant text from the data sources [5]. The *claim matcher* finds all the matching data sources that which mention the same claim in different ways [5]. The *claim checker* verifies them against external information from the internet to compute the accuracy of the claim [5]. The *fact-checker reporter* validates the claims against the facts gathered from *claim matcher* and compiles the accuracy report to publish them through sources like a website, a twitter account, or a slack-bot [5].

In this way, the *ClaimBuster* gives every fact a score between 0 to 1, where a score closer to 1 is more accurate [5]. The model was well trained by using thousands of actual data from general election debates that has been manually fact-checked by humans [5]. The accuracy of the model was measured between 74 and 79 [5]. The system was put to use in real-time for the 2016 U.S. presidential election debates and the results showed a high match between *ClaimBuster* and journalists who checked for the accuracy of the claims [5].

5 CONCLUSION

Big Data coupled with *Artificial Intelligence* and *Machine Learning* can tackle the fact checking problem more efficiently. Rather than working in silos, the social networking giants and the search engine giants should work together with researchers to improve the existing system. This ensures that there are no loose ends with respect to the accuracy of the data. This is important because there

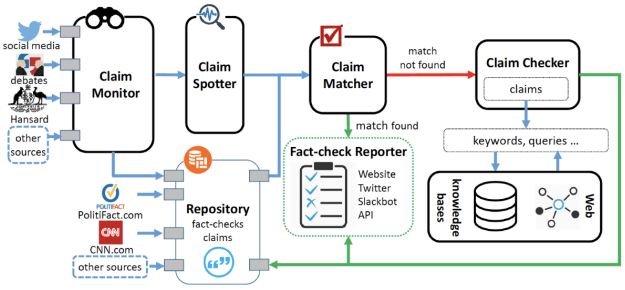

Figure 2: System architecture of ClaimBuster [5]

is always a disconnect between data sources and not everybody has control and access to data that somebody else owns.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Laure Berti-ftquille and Javier Borge-Holthoefer. 2015. *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics.* Morgan & Claypool Publishers, Qatar.

[2] Forbes. 2017. Fake News - Big Data And Artificial Intelligence To The Rescue. Webpage. (Jan. 2017). https://www.forbes.com/sites/jasonbloomberg/2017/01/08/fake-news-big-data-and-artificial-intelligence-to-the-rescue/#69e474df4a30

[3] Forbes. 2017. Fake News: How Big Data And AI Can Help. Webpage. (March 2017). https://www.forbes.com/sites/bernardmarr/2017/03/01/fake-news-how-big-data-and-ai-can-help/2/#7ea468b92039

[4] Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking. In *Proceedings of the 2015 Computation + Journalism Symposium.* Brown Institute of Media Innovation, New York, NY, USA.

[5] Naeemul Hassan, Anil Kumar Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. ClaimBuster: the first-ever end-to-end fact-checking system, In Proceedings of the VLDB Endowment. *Very Large Database Endowment* 10, 1945–1948.

[6] Bernard Marr. 2015. *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.* Wiley, Chichester.

[7] Zhou Zhao, James Cheng, and Wilfred Ng. 2014. Truth Discovery in Data Streams: A Single-Pass Probabilistic Approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14).* ACM, New York, NY, USA, 1589–1598. https://doi.org/10.1145/2661829.2661892

Big Data Applications in the Media and Entertainment Industry

Jiaan Wang
Indiana University Bloomington
3209 E 10th St
Bloomington, Indiana 47408
jervwang@indiana.edu

ABSTRACT

The growth of big data and its various applications in the media and entertainment industry has been swift in recent years as well as the rapid surge of big data and the increasing need for big data technologies. We describe the problems that come with big data and its challenges in the industry. We then present various utilization of big data and why big data is important to the advancement of the media and entertainment industry.

KEYWORDS

i523, hid233, Big data, Media, Entertainment industry, Technology, Recommendation Systems

1 INTRODUCTION

The amount of data being generated is increasing exponentially every year. Currently, we don't have the resources to process or analyze all the data. For example, giant tech companies like Google process over 20 petabytes of data daily [6]. The problem is that data are being generated very quickly and our analytics ability can't catch up with its pace. The goal is to turn these huge amount of data from a burden into decisions and knowledge [2]. Despite that, the technologies used to collect, analyze and interpret data are continuously improving [6].

IDC, the International Data Corporation, believes that companies who take advantage of big data resources to help them make business decisions will flourish while those who don't fully utilize the power of big data will surely lose their competitive edge in the market and find themselves obsolete. This will be especially accurate for companies who face high rates of changes in business [7].

But what exactly is big data? Wanda Group, a multinational conglomerate company based in China, defines big data as a DIKW hierarchical model, which stands for Data, Information, Knowledge and Wisdom [9]. Big data is about the rising challenge that companies face as they handle vast and rapidly-increasing sources of data and knowledge that further introduce a complicated field of inquiries and use issues. Big data technologies define a new era of technologies and frameworks which aim to efficiently gather useful information from huge array of data by using various data science techniques [7].

Emerging sources for big data include industries that are preparing to digitize their content. Particularly, the media and entertainment industry has just started content digitizing five years ago in areas such as "digital recording, production and delivery" [7]. Currently, the industry is gathering "large amounts of rich content and user viewing behaviors" [7]. This will essentially prepare them to adapt for the upcoming big data era and make good use of these so called big data.

2 CHALLENGES IN THE MEDIA AND ENTERTAINMENT INDUSTRY

The issue with the immense data gathering and distributing structure we have created is: "big data is a big mess" [6]. All the information and data captured in our everyday lives simply have nowhere to be processed and it just gets put into storage forever [6].

The media and entertainment industry has always been embracing new technologies because companies believe that big data technologies are crucial to solving their business problems such as reducing operating expenses in an increasingly competitive market and generating enough revenue from producing data and content from various platforms and products [4].

Traditional TV media is facing challenges as its data are scattered. It has physical data which comes from "set-top boxes, network management systems, BOSS systems, etc." [9] as well as online data which comes from user behaviors. The main challenge for traditional TV media in big data applications is data integration [9]. In China, the overall economy of traditional TV media does not look promising. The amount of time user spent on traditional TV has declined while more time is spent on Internet TV. Studies have shown that, "in 2012, Internet TV user base has reached 26.1 million while traditional TV user base is only 600 million" [9]. In addition, "traditional TV operation rate has decreased from 70 percent in 2009 to 30 percent in 2012" [9].

These are the main challenges the media and entertainment industry need to deal with in order to better utilize big data to make a difference:

- Understand different big data sources, whether it is structured or unstructured, such as social media feed, emails, audio and so on. Insights and values can be gained from analyzing these sources to develop better products [4].
- Use complex mathematical algorithms along with domain expertise and information gathering platforms to select and organize information from vast amount of data [4].
- Businesses need to quickly get used to big data. Consumers nowadays are getting more familiar with the idea of big data. In addition, the cost of storage and analytics tools has been greatly reduced. Hence, it is extremely critical for businesses to understand the efficient use of big data to meet their needs and properly set up non-technical aspects such as management of personnel and staff in advance [4].

3 APPLICATIONS IN THE MEDIA AND ENTERTAINMENT INDUSTRY

Enormous amount of information is already being obtained about the entertainment industry [6]. For example, approximately 112 hours of footage has been captured from a horror TV series, *Supernatural* created in 2005 by Eric Kripke with its seventh season in progress. Contained in the footage is information about characters, their actions, dialogues and when, where and how each character dies. These are essentially data that we could use to create a map to the world of *Supernatural* and all its elements. However, currently, all these data are not being used and are stored somewhere, away from us [6].

However, suppose we take all these information and store it in a system. Later, we apply analytics tools using our powerful and growing processing power. Then it is possible that we can create and interact with the world and the characters of *Supernatural* [6].

Big data applications are also widely used in film industries as well. In Indian, a media company teamed up with IBM and ran their predictive modeling algorithm for the movie Ram Leela based on its social media buzz. With the proper selection of cities, the result produced was promising, with a 73 percent success for the movie. In addition, this predictive modeling analysis on social media data was also conducted for movies such as Barfi and Ek Tha Tiger, both of which achieved big success in the film industry. One of the most successful movies in 2013, Chennai Express by Shah Rukh Khan, also used big data analytics techniques supported by *Persistent Systems*, an IT service company, to boost up its social media buzz and marketing strategies. Tweets about Chennai Express generated "over 1 billion cumulative impressions" [3] with "more than 750 thousand" [3] related hash tags in total on twitter over the campaign period of 90 days. Crayon, a big data analytic firm that is based in Singapore, teamed up with producers from leading Hindi film industries to select and release the right kind of music for movies so that they create the perfect hype. Furthermore, Lady Gaga and her associates also used optimization techniques to create the best impact at her live events by going through listening preferences [3].

Another field where big data applications are making influence is sports. Germany, FIFA 2014 champion, has been utilizing SAP's Match Insights software to analyze team performance which made a big difference for the team. It analyzes data such as *touch maps* of player positions, passing abilities, ball retention and so on. In addition, Match Insights was also used by an Indian Premier League team, Kolkata Knight Riders to test the consistency of its players which helped in both auction and ongoing training as well [3].

In order for entertainment companies to create better products and advertising strategies to appeal to more clients, they should figure out why their customers subscribe or unsubscribe using big data analytics [5]. For example, using big data to analyze unstructured data sources such as social media feeds, emails and call records can often reveal reasons, often ignored, for stimulating customer interest. Furthermore, big data analytics also enables the possibility to create personalization systems by combing knowledge from the media and entertainment industry with basic user demographics [5].

One of the most impressive and powerful personalization tools created is recommendation engine. It gathers information from people's past records and suggests or predicts items that they might like. Companies like Amazon has profited by successfully combining its recommendation engine with its online shopping experience from browsing goods to checkout [1].

Artificial intelligence and deep learning technology are just two of the areas that Amazon is spending a vast amount of funds and resources in order to improve its recommendation engine to serve customers more effectively [1]. In May 2016, Amazon announced its complex artificial intelligence platform called *DSSTNE*, pronounced as *destiny*. It is a cloud based open source AI framework created by Amazon to support and boost up its recommendation system [1].

In addition, Amazon claims that customers frequently watch and review pilots created and released at Amazon Studios. Executives from Amazon then choose which pilots will develop into a full series based on customer feedback. *Transparent*, a comedy show based on a transgender patriarch whose family lives in Los Angles, was one result from this system. After its initial release in 2014, it received positive reviews from the general public and raised awareness about transgender problems [8]. It was rewarded the following year with "the Golden Globe for best TV series, musical or comedy" [8].

Another big media company who uses recommendation engines big time is Netflix who believes that content discovery is the number one priority. This is not surprising because its on-demand video and media streaming possibly dominates the world's market for digital content consumption. Just like Amazon, Netflix has spent immense amount of expenses and resources to make sure that its recommendation engine is one of the best to display its content library as much as possible [1]. In December 2015, Netflix remodeled its recommendation system, "deciding to do away with region based preferences" [1] due to their continue international expansion.

Netflix utilizes sophisticated statistical equations to advertise series that customers might like, for example, *House of Cards* and *Orange Is the New Black*. Those equations usually contains predictor variables such as popularity of the series, customers' previously watched content or their demographics [8].

The personalization systems that Netflix has created for its customers has helped to lower the number of subscription cancellations, saving more than 1 billion dollar a year [8].

4 CONCLUSION

The rapid growth of big data has given the media and entertainment industry an unique opportunity to utilize resources in order to benefit from big data applications and technologies. However, there are still some key challenges media companies are facing such as how to quickly adapt to the big data era, how to deal with and analyze immense amount of data pouring in every minute and how to make cost-effective products and consumer experiences. Examples of current effective big data applications and technologies such as Match Insights from SAP and personalized recommendation engines from Amazon and Netflix are provided. In summary, big data applications and technologies are crucial in the success of media and entertainment companies.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Shabana Arora. 2016. Recommendation Engines: How Amazon and Netflix Are Winning the Personalization Battle. Web Page. (June 2016). https://www.martechadvisor.com/articles/customer-experience/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/ HID: 233, Accessed: 2017-10-06.

[2] Lauren Browning. 2015. We sent men to the moon in 1969 on a tiny fraction of the data that's in the average laptop. Web Page. (June 2015). http://www.businessinsider.com/mind-blowing-growth-and-power-of-big-data-2015-6 HID: 233, Accessed: 2017-10-07.

[3] Ashok Karania. 2014. How Big Data Is Changing The Entertainment Industry! Web Page. (July 2014). https://www.linkedin.com/pulse/20140730194648-8949539-how-big-data-is-changing-the-entertainment-industry HID: 233, Accessed: 2017-10-03.

[4] Helen Lippell. 2016. *Big Data in the Media and Entertainment Sectors* (1 ed.). Springer International Publishing, Gewerbestrasse 11 CH-6330 Cham (ZG) Switzerland, Chapter 14, 245–259. https://doi.org/10.1007/978-3-319-21569-3_14 HID: 233, Accessed: 2017-10-03.

[5] Ritesh Mehta. 2017. Big Data in the Field of Entertainment. Web Page. (Aug. 2017). https://insidebigdata.com/2017/08/20/big-data-field-entertainment/ HID: 233, Accessed: 2017-10-03.

[6] Tawny Schlieski and Brian David Johnson. 2012. Entertainment in the Age of Big Data. *Proc. IEEE* 100, Special Centennial Issue (May 2012), 1404–1408. https://doi.org/10.1109/JPROC.2012.2189918 HID: 233, Accessed: 2017-09-20.

[7] Richard L. Villars, Carl W. Olofson, and Matthew Eastwood. 2011. Big data: What it is and why you should care. *White Paper, IDC* 14 (June 2011). www.tracemyflows.com/uploads/big_data/idc_amd_big_data_whitepaper.pdf HID: 233, Accessed: 2017-09-20.

[8] Angus Whitley. 2016. How Entertainment Companies Use Big Data. Web Page. (July 2016). https://www.comstocksmag.com/bloomberg/how-entertainment-companies-use-big-data HID: 233, Accessed: 2017-10-03.

[9] Chunjie Zhang, Wenqian Shang, Weiguo Lin, Yongan Li, and Rui Tan. 2017. Opportunities and challenges of TV media in the big data era. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, Wuhan, China, 551–553. https://doi.org/10.1109/ICIS.2017.7960053 HID: 233, Accessed: 2017-09-20.

Recommendation Systems on the Web

Jordan Simmons
Indiana University Bloomington
jomsimm@iu.edu

ABSTRACT

Recommendation Systems are being used all over the web. There are different popular techniques that are being used in modern systems. Some of the larger well know companies are using this technology very well. To better understand Recommendation Systems, we provide an overview of some techniques, state of the art systems, and challenges and limitations of Recommendation Systems.

KEYWORDS

i523, hid336, Recommendation Systems, Big Data

1 INTRODUCTION

Recommendation Systems (RS) leverage big data in order to create value for both businesses and customers. RS can be understood as systems that "generate meaningful recommendations to a collection of users for items or products that might interest them. " [7]. RS are effective for a variety of industries and products which can range from a product in a store, a news article on a site, or a search query. RS is beneficial to businesses and customers by increasing metrics such as revenue and customer satisfaction [2]. Many online platforms are starting to use RS to analyze their data. In order to gain a better understanding of RS, general analysis of modern techniques, companies currently using RS, and challenges and limitations within the field will be covered.

2 RECOMMENDATION TECHNIQUES

Three common RS techniques would include content-based, collaborative, and hybrid recommendations [1]. Other techniques exist, but these three are the most widely used today. In order to determine which technique is best depends on the recommendations to be made, and the data used to make them. Many times, the hybrid approach is used because there can be limitations with other approaches [1]. Overall, it is best to understand a little bit about each technique before choosing which is best.

2.1 Content-Based

Content-based RS recommend items to users by using descriptions of items and how the user is profiled based on their interest [8]. Items are classified by different characteristics, attributes, or variables [8]. Once items are classified, they can be grouped together based on the classifications. Users are classified by data they provide to the system, and/or the data collected by interacting with the system.

Content-based RS are commonly seen on web applications and E-commerce sites. These types of systems readily track and monitor almost all user activities. Typically a user has an account with the system, which is where data was voluntarily provided. With this data, users can be classified easier compared to a customer walking into a brick and mortar business.

2.2 Collaborative Filtering

Collaborative filtering can be described as the "process of filtering or evaluating items using the opinions of other people" [10]. This type of RS is commonly seen on systems where an item can be rated by a user. With this technique, user rating are collected and stored from a user for an item that they have used or purchased. The ratings from the user are then compared to other users that have rated the same item. For example, person A buys items 1 and 2 and rates each item highly. Then, person B buys item 1 and rates it highly. Since person A and B both bought and rated item 1 highly, the system would likely recommend item 2 to person B. On the contrary, if person B gave item 1 a low rating, the system would not likely recommend item 2 to person B. This concept uses the assumption that "people with similar tastes will rate things similarly" [10]. This assumption may not be true in all cases, but it is a good base for RS to start learning users interests, and recommend items based on those interest. With this technique, the more ratings that the systems has collected per item, and the more ratings given by the user, the easier it is for that system to make recommendations to that specific user.

2.3 Hybrid

Hybrid RS takes two or more techniques and combines them to improve performance and reduce limitations that a single technique might have [3]. In most cases, collaborative filtering is used with one or more of the other techniques to improve performance. Other techniques that are used and not discussed include demographic, utility-based, and knowledge-based recommendations [3]. The hybrid approach narrows down items with one technique, and then uses another technique on that subset of items to make a more accurate recommendation. Determining the best hybrid system depends on the specific business case, and the data used to make the recommendation.

An example of a hybrid approach would use collaborative filtering and the content-based methods described above. For example, if User A is interested in baseball. The system would use the content-based approach to narrow down all items that are classified as baseball items. From this subset of baseball items, the system could then use the collaborative-filtering approach to find the items with ratings from other users which will be user group B. The system would then find all item ratings from user group B and compare those item ratings to person A. If there are any users in group B that have similar likes to person A, the system would likely recommend the baseball items to person A that person B has previously rated highly. This is a high-level example of how a hybrid RS would work. Real world examples are more complex than this example, and use large amounts of data.

3 MODERN SYSTEMS

Two well known companies that are currently using RS are Netflix and Amazon. These two companies have huge customer bases, in which they collect data on. The data collected within these sites and how they utilize it to generate suggestions to their users is what makes these companies have successful advanced recommendation systems.

3.1 Netflix

Netflix is an internet based company that offers a variety of movies and television shows. Netflix had a problem of customers sorting through its large selection of movies and shows, and eventually losing interest which resulted in abandonment of their services [5]. Over the years, Netflix has created and continually developed new RS algorithms which they claim saves them more than one billion dollars per year and a monthly turnover in the low double digits [5].

Netflix does very well at recommending movies and shows to its users. They have incorporated different strategies to collect data from users which is the base of their RS. Data is collected in the form of customized search, video ratings, continue watching feature, amount of time spent watching and other user activities [5]. Using the data collected from these features, Netflix can recommend top rated, now trending, and videos based on user interest, which is very appealing to the user when there are so many selections to choose from.

3.2 Amazon

Amazon is an online store that sell a large variety of products. Amazons RS provides recommendations for millions of customers from a catalog that has millions of products. [11]. Instead of comparing customers to customers, amazon uses an item-based collaborative filtering approach. This process finds items that were bought together with unusually high frequencies, and uses these relationships to recommend products to customers based on what they have purchased in the past [11]. With this algorithm, Amazon is providing a unique experience to every user and helping them find products they may not have found. Since the initial launch of this algorithm, it has been adjusted to make it easier for people to find material, compared to other algorithms, and adapted to help solve many other problems. [11]

4 CHALLENGES AND LIMITATIONS

As with most technologies, RS has its challenges and limitations. It is hard to speak of this topic without speaking about the questions "more data usually beats better algorithms" [9]. This quote has raised controversy about which of the two actually produce better results. In most cases, there are many different variables to consider when answering this question.

4.1 Limitations

With complex systems, there can be many variables that cause issues that limit full capabilities of that system. Specifically, in RS, some of these limitations include cold start problems, data sparsity, limited content analysis, and latency problems [6]. These limitations seem to be more data related rather than the actual techniques and approaches of the technology being used to analyze that data. When there is no data for a new user, it is hard for RS to create suggestions for this user. The system has no data on the users activities or what interests that user has. When a new item is added to a system, there are no reviews and no data collected with the interaction of user for this particular item. On the other hand, too much data can become redundant. At this point gathering more data will have limited gains.

4.2 Cross-Domain Recommendations

Cross-domain RS aim to "leverage all the available user data provided in various systems" and domains" which allows systems to "generate more encompassing user models and better recommendations" [4]. Every day the amount of data being collected increases. This data is being collected from different sources. Cross-Domain RS could use data from different sources, which could make up for some of the data caused problems. An example of a cross-domain recommendation would be Netflix using data from Facebook to help recommend movies to a new user. Using data from various systems like this would bring up new issues like privacy and security, but if systems started working together and sharing data there could be benefits for both systems.

Cross-domain recommendations help with domain specific data issues. Two different systems may have different ways of collecting and organizing data. If system 1 collects variables A , B and C, and system 2 collects variables A, B, and D, each system has information that the other system does not have. This is where sharing the data between systems could have benefits for both systems. In doing this, each system is not only benefiting from more data, but different and perhaps better data. This would also require using better algorithms to analyze the different sets of data. Depending on the system, more data can be more beneficial than better algorithms. In terms of scaleability, gathering more data that is different from what is currently being collected, and using better algorithms along with the different data could potentially maximize recommendations for that system.

5 CONCLUSION

With a base understanding of RS, it is easy to see how this technology can be very beneficial in online platforms. RS has different techniques that can be used in a variety of online systems. Many large companies are creating custom RS and are benefiting greatly from them. As the massive amount of data grows from day to day, the ways in which RS is used will continue to evolve. It will be interesting to see how cross-domain recommendations are used in the future, and if companies start to adopt this concept of sharing data. Data being analyzed from various systems could unlock hidden information that a single system may not be capable of producing.

ACKNOWLEDGMENTS

The author would like to thank course instructors for organizing setup of the latex format.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17, 6 (June 2005), 734–749. https://doi.org/10.1109/TKDE.2005.99

[2] Xavier Amatriain and Justin Basilico. 2016. Past, Present, and Future of Recommender Systems: An Industry Perspective. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 211–214. https://doi.org/10.1145/2959100.2959144

[3] Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction* 12, 4 (01 Nov 2002), 331–370. https://doi.org/10.1023/A:1021240730564

[4] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. 2015. *Cross-Domain Recommender Systems*. Springer US, Boston, MA, 919–959. https://doi.org/10.1007/978-1-4899-7637-6_27

[5] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (Dec. 2015), 19 pages. https://doi.org/10.1145/2843948

[6] Shah Khusro, Zafar Ali, and Irfan Ullah. 2016. *Recommender Systems: Issues, Challenges, and Research Opportunities*. Springer Singapore, Singapore, 1179–1189. https://doi.org/10.1007/978-981-10-0557-2_112

[7] Prem Melville and Vikas Sindhwani. 2010. *Recommender Systems*. Springer US, Boston, MA, 829–838. https://doi.org/10.1007/978-0-387-30164-8_705

[8] Michael J. Pazzani and Daniel Billsus. 2007. *Content-Based Recommendation Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 325–341. https://doi.org/10.1007/978-3-540-72079-9_10

[9] Anand Rajaraman. 2008. More Data Usually Beats Better Algorithms. (03 2008). http://anand.typepad.com/datawocky/2008/03/more-data-usual.html

[10] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 291–324. https://doi.org/10.1007/978-3-540-72079-9_9

[11] Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.Com. *IEEE Internet Computing* 21, 3 (May 2017), 12–18. https://doi.org/10.1109/MIC.2017.72

Big Data Analytics for Research Libraries and Archives

Timothy A. Thompson
Indiana University Bloomington
School of Informatics, Computing, and Engineering
Bloomington, Indiana 47408
timathom@indiana.edu

ABSTRACT

Research libraries and archives have played a longstanding role in information management and access. In the second half of the twentieth century, libraries were at the forefront of automation and networked access to information. Since the advent of the internet, however, they have failed to keep pace with technological advances and currently face challenges in serving the evolving needs of researchers, whose information-seeking strategies are now shaped by internet search engines and online social media applications. To remain relevant in the current information landscape, libraries and archives must implement new strategies for converting legacy metadata to new formats that can add value to the research process. Although the data and metadata produced by libraries and archives may not always qualify, *prima facie*, as big data, an awareness among information professionals of the tools, techniques, and affordances of big data analytics can help make library services more relevant to researchers.

KEYWORDS

i523, HID340, Library Metadata, Archival Metadata, Linked Open Data, Data Conversion

1 INTRODUCTION

Cultural heritage institutions such as libraries and archives have a longstanding tradition of producing structured data—in the form of catalog records or finding aids—to describe their collections. In the twentieth century, library card catalogs were gradually replaced by machine-readable formats, the foremost of which were the MAchine Readable Cataloging (MARC) formats for bibliographic and authority data (standardized as ISO 2709 and ANSI/NISO Z39.2) [3].

Initial development of the core MARC format, commissioned by the Library of Congress, was finalized in 1968, when the first electronic catalog records were distributed [3]. Originally, MARC records were used to facilitate the automated creation of card catalogs, which remained the primary method of information retrieval in libraries until the 1980s, when online public access catalogs (OPACs) became available and the pace of automation began to accelerate [9]. It was not until 2004 that the MARC record format (stored as a binary file) was mapped to an XML schema, making it more amenable to computation and transformation [7]. Notwithstanding, the MARC format has now become increasingly archaic and hinders data sharing and interoperability between libraries and contemporary platforms for research and information retrieval, such as the World Wide Web.

2 IS LIBRARY METADATA BIG DATA?

Although libraries and other cultural heritage institutions have created millions of metadata records over time, even the largest library catalogs fall short of the scale typically associated with big data. The entire catalog of the Library of Congress, an institution that holds over 13 million physical volumes, totals less than 100 gigabytes. By comparison, Twitter produces approximately 12 terabytes of data on a daily basis [2, p. 1527]. If the definition of big data limited only to measurements of storage, then approaching library data in terms of big data would seem to be excessive [9]. However, if big data is defined more broadly as a set of methodologies for analysis and an ecosystem for data aggregation, then libraries clearly stand to benefit from adopting its tools and techniques.

In one view, big data constitutes a "social movement," shaped by alliances "among heterogeneous players in business, academia, and government" [2, p. 1527]. By undertaking projects focused on data modeling and mass conversion and migration of legacy data, libraries can position themselves to partner with other players and provide enhanced information retrieval services, exposing their metadata in contexts that are more relevant to the current needs of researchers. In addition, by adopting graph-based models that are native to the World Wide Web, libraries can merge their data more seamlessly with the wider universe of online data in order to "generate massive collections of new relationship assertions" [9].

By leveraging universal standards such as the Resource Description Framework (RDF), libraries, archives, and other cultural heritage institutions can uncover latent relationships that are currently buried in catalog records, connecting them to data from disparate sources and providing a "training set for all human knowledge" [9, p. 430]. Teets and Goldner suggest that the process of splitting catalog records into discrete, linkable statements could vastly expand the size and scope of library-created metadata: multiple data points could be extracted from a single catalog record, enabling the creation of new datasets around authorship networks or histories of publication, for example. This newly minted data could be further enhanced with linkages to data sources that exist outside of the library domain—including Wikidata, open government data, and digital or digitized texts published and hosted by nonprofit organizations such as the Internet Archive (https://archive.org) [9].

3 TOWARD BIG DATA

Both libraries and archives face particular challenges in attempting to embrace the ethos of big and complex data. The rules and instructions used by catalogers and archivists to describe information-bearing resources are still reflective of the card catalog environment and do not support the kind of data-centric granularity needed to enable effective data integration and interoperability [10]. One of

the primary obstacles in converting and migrating legacy data is the problem of entity resolution and name disambiguation. A second obstacle, one that is by turns social, legal, and technical in nature, involves libraries' ability to publish and preserve digitized content.

3.1 Entity Resolution

Two recent projects exemplify the large-scale effort in libraries and archives to remediate legacy data and merge information from multiple sources. In the archival community, researchers are often faced with scenarios in which a person's papers are scattered among geographically distant repositories, but there is no master index that links the relevant collections together. One initiative, the Social Networks and Archival Context Project (SNAC) (http://snaccooperative.org/), is working to develop algorithms and routines for entity resolution in order to address this problem. Researchers in the SNAC Project have focused on developing supervised machine learning algorithms for matching names across records that have been collected from multiple archival repositories [6]. Experiments with methods based on Naive Bayes Classification have yielded promising results, particularly when data from name strings is combined with contextual information (such as birth and death dates) that has been extracted from related records, with an accuracy rate of approximately 80% [6].

The task of entity resolution is made particularly difficult by the approach to data creation that has been traditionally employed by libraries and archives. In catalog records describing authors, books, or archival collections, creators are identified by name strings rather than unique identifiers. Catalogers must follow detailed rules for ensuring that each name string—known as an "authorized heading"—is unique, but because these strings are hand-crafted by humans rather than generated by machines, they are particularly vulnerable to error and inconsistency.

In a global database such as Wikidata (https://www.wikidata.org), by contrast, which was originally compiled from structured data templates on Wikipedia pages, the American author Mark Twain is represented by a unique identifier that can be dereferenced as an HTTP URI: https://www.wikidata.org/wiki/Q7245. Wikipedia pages about Twain, regardless of the language they are written in, are able to link to this single identifier. In the Library of Congress Name Authority File, however, Twain is identified instead by the string "Twain, Mark, 1835-1910."

The Library of Congress maintains the "authorized" list of names used by U.S. libraries as controlled access points, but many other national libraries also maintain their own authority files. In the case of a well-known author such as Twain, there may be substantial agreement across institutions from Roman-script language communities as to the format of the authorized heading. For libraries and archives whose official languages are expressed in other character sets, the process of entity resolution may be more difficult.

To address the problem of string-based identification, the OCLC Online Computer Library Center, a global data provider for the library industry, has developed an initiative called the Virtual International Authority File (VIAF) (http://viaf.org/). VIAF is a data aggregation portal that attempts to resolve named entities across "more than 130 million authority and bibliographic records expressed in multiple languages, scripts, and formats" [4, p. 1]. MARC authority records are contributed to VIAF from nearly 50 contributing partners, most of which are national libraries [4]. Across this corpus of records, the VIAF project clusters and merges references to named entities using a 300+ core Hadoop cluster in a monthly batch process that takes approximately "12 hours of cluster compute time to complete" [4, p. 2]. In a multistep algorithm, named entities in VIAF are progressively grouped into identity clusters, and pair-wise matching is performed between datasets from each institutional contributor. Because it is able to draw on a wider range of sources for disambiguation and entity resolution, VIAF has, to date, achieved a higher degree of accuracy in matching entities than has the SNAC Project, with success rates of over 90% [4].

3.2 HathiTrust

In the library domain, the project that perhaps comes closest to the scale and scope of "big data" is the HathiTrust Digital Library and the related HathiTrust Research Center. In large part, the HathiTrust initiative grew out of the response of major research libraries to the Google Books mass digitization enterprise [1, 8, 11]. Libraries were particularly concerned about the issues of long-term digital preservation and open access to research data. With the favorable settlement of a high-profile lawsuit brought by the Authors Guild and other plaintiffs against both Google and HathiTrust, the latter has moved ahead with research projects to publish curated datasets extracted from the full text of both public domain and in-copyright titles. HathiTrust is committed to providing "non-consumptive" access to its data, and it has developed an approach that provides access through "data capsules"; this approach gives researchers as much flexibility as possible while simultaneously protecting against the unlawful "leakage" of full-text content onto the open web [11].

Through HathiTrust, researchers are now able to perform data mining on an extracted features dataset that contains page-level data features from all of the nearly 14-million volumes in the HathiTrust corpus. Although this dataset is substantially larger than the largest catalog of library metadata, its current size of 4 terabytes is still comparatively small by big data standards [5]. Nonetheless, HathiTrust's methodological sophistication and principled approach to data usage and access provide a model for other projects in the library domain to follow.

4 CONCLUSION

For libraries, archives, and other cultural heritage institutions, the most significant paradigm shift that could be attributed to the big data phenomenon is a new view of descriptive metadata *as* data in its own right. As libraries in particular move away from legacy formats and domain-specific idiosyncrasies, they will be better equipped to serve the evolving needs and interests of researchers, who may themselves be struggling to come to grips with the scale of data in the age of the internet. Once libraries gain a more sophisticated understanding of their own data models and formats, they will be better positioned to assist researchers in managing, storing, and sharing their data—which is likely to be much bigger than anything produced by libraries themselves.

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] H. Christenson. 2010. HathiTrust: A Research Library at Web Scale. *LRTS* 55, 2 (2010), 93–102.

[2] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto. 2015. Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology* 66, 8 (2015), 1523–1545.

[3] K. M. Ford. 2012. LC's Bibliographic Framework Initiative and the Attractiveness of Linked Data. *ISQ: Information Standards Quarterly* 24, 2/3 (2012), 46–50. http://www.niso.org/publications/isq/2012/v24no2-3/ford/

[4] T. B. Hickey and J. A. Toves. 2014. Managing Ambiguity in VIAF. *D-Lib Magazine* 20, 7/8 (2014), 1–12.

[5] A. Kinnaman and E. Dickson. 2017. HTRC Docs: Extracted Features Dataset. (Sept. 2017). https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset accessed 2017.

[6] R. R. Larson and K. Janakiraman. 2011. Connecting Archival Collections: The Social Networks and Archival Context Project. In *Research and Advanced Technology for Digital Libraries, TPDL 2011*. Springer, Berlin, 3–14. https://doi.org/10.1007/978-3-642-24469-8-3

[7] Library of Congress. 2004. MARC XML Design Considerations. (Dec. 2004). http://www.loc.gov/standards/marcxml/marcxml-design.html accessed 2017.

[8] B. Plale, R. McDonald, Y. Sun, I. Kouper, R. Cobine, J. S. Downie, B. Sandore Namachchivaya, and J. Unsworth. 2013. HathiTrust Research Center: Computational Access for Digital Humanities Research and Beyond. In *JCDL'13*. Association for Computing Machinery, Indianapolis, Indiana, USA, 395–396.

[9] M. Teets and M. Goldner. 2013. Libraries' Role in Curating and Exposing Big Data. *Future Internet* 5 (2013), 429–438. https://doi.org/10.3390/fi5030429

[10] R. Tennant. 2002. MARC Must Die. (Oct. 2002). http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/ accessed 2017.

[11] J. Zeng, G. Ruan, A. Crowell, A. Prakash, and B. Plale. 2014. Cloud Computing Data Capsules for Non-Consumptive Use of Texts. In *ScienceCloud 2014*. Association for Computing Machinery, Vancouver, BC, Canada, 9–15.

Big Data Dangers: Weaponizing Social Media

Ross Wood
rmw@indiana.edu
HID 345

ABSTRACT

Social media has changed the way people get information, disseminate information, communicate, and stay in touch with others, both online and in the real world. As more and more people from different age groups and socioeconomic backgrounds begin adopting social media and becoming active users, data is being created at a geometric rate. The analysis of all this data being generated can be used in a myriad of different ways, including nefarious ones. It is possible to analyze the digital footprint of social media users in order to accurately target enormous swaths of a population with propaganda, misinformation, and deception which have been created to cater to the specific population's social or political bias.

KEYWORDS

i523, HID345, Social Media, Social Media Mining, Big Data, Social Media Scraping

1 INTRODUCTION

The rise of social media among all tiers of society, not just the tech savvy portion, has created interesting opportunities in the field of big data and social media mining. The decrease in costs and size of computing tools is also helping to fuel a technological explosion among different societies, with more and more people having access to social media than ever before. This increase in users creates a tremendous amount of data, all of which can be analyzed to reveal information about the users and real world populations. This information can be used to inform, educate, and improve the lives and systems we use daily. However, in the wrong hands, this kind of information can also be used to influence and manipulate citizens into supporting things that are against their self-interest, and against the interests of their society. If an informed citizenry is essential to maintaining freedoms, then in effect, social media can be weaponized and used to curtail freedoms for some by misinforming and radicalizing its user bases.

2 USER BASE EXPLOSION

The increase in population of social media user bases is helping lead the way in 21st century social engineering, for better or worse, and this increase is causing data to be created at a scale that has never before been seen. Indeed, a report found that the population of adults in the United States who use social media rose from 7% in 2005, to 65% in 2015 [6]. Furthermore, the report found that "there continues to be growth in social media usages among some groups that were not among the earliest adopters, including older Americans" [6]. The ability to scrape massive amounts of user data from social media sites allows for an analysis of a user's individual digital footprint. When all these footprints are put together and analyzed, conclusions about an individual's taste in entertainment, political, and social leanings can be drawn, as well as information about an individuals personality and socioeconomic background. When these conclusions are combined with user location and network structure data, a situation is created where an organization or group could manipulate an entire segment of a country's population. The success of this method depends largely on how accurate all the accumulated user data is. The more accurate the data, the better job a machine does at making these demographic predictions. But just how accurate can a machine be at predicting a human's personality and sociopolitical leanings based on digital information alone?

2.1 Accuracy

The accuracy of a machine's prediction about individual personality and demographics is improved as it accumulates more user data to work with. In essence, the more a person uses social media and creates information about themselves, the easier it is going to be for a machine to look at this information and predict certain things about the person. One study found that to a certain point, humans are better than machines at making personality judgments on other humans. However, once a machine has, in this example as little as 100 Facebook likes, the machine's ability to make accurate personality judgments starts to outperform the predictive ability of humans [9]. The study found that with even a small amount of data, machines can predict a person's personality better than that same person's close acquaintance. The study's findings also "highlight that people's personalities can be predicted automatically and without involving human social-cognitive skills" [9]. This automated process makes it easy for people and organizations to gather large amounts of user data for analysis, that can then be used however these people or organizations see fit.

3 NEFARIOUS DEMOGRAPHIC INFORMATION ANALYSIS

An individual's personality traits are not the only information that machines can gleam from peoples' social media footprints. Indeed, there are a number of different methods of analysis that can be used when approaching this problem. One would use different methods for trying to figure out different kinds of population information during analysis. It is possible to infer user data, such as age, race, and socioeconomic background, based only on concepts as simple as word use and assigning emotional feelings to different words [7]. All of this demographic information would be quite valuable to individuals or organizations wanting to understand or influence various populations and subsets of populations, or even just to understand user bases for marketing purposes. This user population information that has been generated could then be used to target these various population segments with misinformation and propaganda that appeals to their specific confirmation bias. If the misinformed user were to then share the misinformation or propaganda with their like minded friends online, this could create a self-sustaining cycle of misinformation that reinforces the

misinformed beliefs of users. This is known as an echo chamber and they can be quite pervasive in social media [2].

3.1 Using Demographic Data

There are recent examples of these techniques having been used on populations. A post electoral analysis of various elections in Russia found that not only are governments using this approach to their benefit, but also that it works best in regions that exhibit large amounts of racial, social, religious, or socioeconomic tension [4]. Using this data to spread misinformation works by causing both sides of any argument to seem radical, even the rational side. The misinformation does this by working off the confirmation bias of the reader, which was acquired by analyzing their digital footprint. The example of Russian election use found that this approach worked best in areas that were particularly volatile in regards to racial prejudice and struggle [4]. So in other words, someone with a racial bias towards a certain group would have their beliefs reinforced through social media use. This effect is only further reinforced by social media users whose experience on social media is limited primarily to echo chambers, which further distort their view of society while simultaneously widening the societal divide and radicalizing users [2].

3.2 Bots

Misinformation and propaganda, which add fuel to the fire of discussions in these so called echo chambers, can also be spread by the insidious use of bots, which are programs that do automated tasks. These tasks range from simple jobs like retweeting something, to complex tasks like conversing with a human and tricking them into thinking the bot is real. Whatever the case, that task is often one that helps convince people of an agenda that the bots have been told to push. As of March 2017, there are almost 48 million active twitter bots. These bots can push any information their masters want, while also serving the purpose of inflating the popularity of people, tweets, and points of view that are more aligned with the bots' agenda [1]. Bots are key to the successful spread and propagation of misinformation and their campaigns. They are now capable of even greater insidiousness by being able to target specific groups using user data generated by social media users. This misinformation technique is taken to another level with bots sophisticated enough to engage in limited conversations with real people over social media [8]. Chat rooms, message boards, and social media sites are flooded with bots, all pushing different agendas. The more resources and effort an organization can put behind an misinformation push of this manner, the more more effective it will be.

3.3 Social Polarization

Irregardless of who uses these techniques and for what purposes, one outcome that always arises from their use in this way is social polarization. The focal point of the polarization depends on which social struggle is being exploited to manipulate the common social media user. Protests, economic inequality, racial discrimination: there are any number of current social problems to draw from if one wanted to fan the flames of social unrest on a large scale in order to push a political or corporate agenda. This effect is beginning to have a visible influence on societies around the world as misinformation campaigns are causing more and more people to be misinformed on current events. A misinformed voter does not make good choices, and enough of them together has the potential to throw the entire democratic process out of whack [3]. Social polarization leads to instability and unrest, which can be profitable to certain members of society who might take advantage of these techniques.

4 DISCUSSION

Social media use has become so ingrained into everyday life that, at this point, it would be almost impossible to get people to stop using it, even if it was demonstrated to them that it has a potential negative effect on society. Indeed, this abstinence approach should not be advocated, as social media and the data it produces can be used to benefit society at large in a multitude of ways. That being said, it would be wise to continue to monitor and study different ways social media can harm society by being used to benefit the few at the expense of the many [5]. If processes and protections aren't put into place in the near future, the entire democratic process could continue to destabilize and become polarized to the point where it is so consumed by corruption that it cannot be salvaged.

One possible solution is to examine your network and establish who the key peddlers of misinformation are and to block or delete their accounts. This approach would be effective since "successful sources of false and biased claims are heavily supported by social bots" [8]. Another approach is to fight fire with fire and create bots that are sophisticated enough to detect misinformation as it begins to trend and then counter this trend with the truth [1]. Whatever solution to this problem takes shape, the exponential growth of social media users and the unprotected data they generate [6] make it imperative that a solution is found and implemented. Until that happens, huge portions of social media users are going to continue to be tricked into believing misinformation and propaganda through data analysis and manipulation.

5 CONCLUSION

The rising population of social media users is beginning to pose a threat in regards to a population's ability to stay accurately informed. As this population of users grows and creates more and more data, so to does the ability to use sophisticated techniques to deceive the users. The growth of social media grows hand in hand with new dangers. As more people get their news through social media, it becomes easier to misinform them. In essence, the more information that is known about someone, the easier it is to take advantage of them. And if you are trying to dupe someone, the modern world makes it easy to accumulate information on people by analyzing their social media digital footprint.

We are starting to see real world effects of these techniques in the form of population destabilization and user manipulation through propaganda and misinformation campaigns. At present there are no safeguards in place to protect users from attempts to deceive them, no matter where the attacks come from or what agenda they have. Until safeguards are developed and put into place to protect users and the enormous amounts of data that they generate from those who would use it against them, the problem is only going to continue to grow and get worse.

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] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. *Limiting the Spread of Misinformation in Social Networks.* ACM, New York, NY, USA, 665–674. https://doi.org/10.1145/1963405.1963499

[2] Siying Du and Steve Gregory. 2017. The Echo Chamber Effect in Twitter: does community polarization increase?. In *Complex Networks and Their Applications V*, Vol. 693. Springer, Cham, Cham, Switzerland, 373–378.

[3] Robert Epstein. 2016. *Subtle New Forms of Internet Influence Are Putting Democracy at Risk Worldwide.* Springer New York, New York, NY, 253–259. https://doi.org/10.1007/978-1-4939-6415-4_9

[4] Regina Goodnow, Robert Moser, and Tony Smith. 2014. Ethnicity and Electoral Manipulation in Russia. *Electoral Studies* 36 (12 2014), 15 – 27.

[5] Rodrigo Ochigame and James Holston. 2016. Filtering Dissent Social Media and Land Struggles in Brazil. *New Left Review* 99 (Jan. 2016), 85 – 100. https://newleftreview.org/II/99/rodrigo-ochigame-james-holston-filtering-dissent

[6] A. Perrin. 2015. *Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites–a Nearly Tenfold Jump in the Past Decade.* Pew Research Center, Washington D.C. https://books.google.com/books?id=OupAnQAACAAJ

[7] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLOS ONE* 10, 9 (sep 2015), e0138717. https://doi.org/10.1371/journal.pone.0138717

[8] 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

[9] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (jan 2015), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Big Data Applications in Astronomy and Astrophysics

Ricky Carmickle
Indiana University
901 E 10th St
Bloomington, Indiana 47408, USA
rcarmick@umail.iu.edu

ABSTRACT

This paper will provide an overview of how Big Data has been applied to existing astrophysics and astronomical datasets.

It will discuss the two-way relationship between big data and astrophysics which has seen each field advance because of the pressures applied and capabilities developed by the other.

Improvements in cameras, which capture the observations from telescopes, have opened the door to more detailed imagery of the sky which can be generated at increasing rates and observation platforms are harnessing this detail into more precise astronomical research projects which challenge every part of the big data pipeline

KEYWORDS

i523, hid304, Big Data, Astronomy, Astrophysics, Astroinformatics, Astrostatistics, Large Synoptic Survey Telescope, Square Kilometer Array, Sloan Digital Sky Survey, SkyMapper Southern Sky Survey, Hubble Space Telescope, James Webb Space Telescope, Hough Transform, Renewal String,

1 INTRODUCTION

The volume of data generated by astrophysics and astronomical platforms rivals the output of other data sources. Astrophysics and astronomy are considered a primary domain generating 'Big Data', alongside Twitter, YouTube, and Genomics research[25]. The growth in volume of astronomical data has been driven by improvements to camera technology, in the sensitivity of sensors, and in the computational resources to gather and store imaging data gathered by astronomy research projects. These are closely related and are now largely collaborative. Universities and researchers from across countries will collaborate with space programs and privatized researchers to create observational platforms which can fulfill the needs of all parties involved.

Astronomy's transition into big data began with the release of the Hubble Deep Field (HDF) image in 1995[2]. The objects in the HDF were cataloged, and for the first time astronomers accessed a searchable database to return data on objects related to their work rather than the raw images.

2 BIG DATA IN ASTROPHYSICS

Astronomic data requires perpetual development of data cleaning, storage, processing, searching, mining, and analysis tools[5].

The foremost challenge is development of search and query tools which can return data relevant to the needs of particular researchers or institutions who participated in the creation of a given sky survey or observational platform. The importance of search functionality is due to the collaborative nature of Astronomical Big Data. The largest observational platforms are expensive in both computing needs and hardware, which incentivizes institutions and agencies to collaborate on projects so that the different needs of astronomers, whatever their particular field of study, can be addressed by searchable databases. Researchers typically focus on particular astronomical phenomena, such as quasars, galaxy formation, exoplanets, etc, and must retrieve data on their specific study area relatively quickly.

The data collection tool for astrophysics and astronomy data is the telescope. The telescopes applicable to big data, and those which astrophysics research relies on, are either space telescopes placed in orbit or larger terrestrial telescopes. These observe astronomical objects in high definition over a wide range of the electromagnetic spectrum from gamma rays, to visible light, to extremely low-frequency radio waves[23].

The largest astronomic and astrophysical research projects have created databases of hundreds of terabytes. Survey platforms in development are expected to capture data in the exabytes[16, 22, 24].

The Large Synoptic Survey Telescope (LSST) is the highest volume astronomical data project currently under construction. The LSST is expected to generate 15 terabytes of data per day with over 200 dimensions of data per astronomical object[21]. This 3.2 gigapixel telescope camera is located in Cerro Pachn, Chile, and is expected to generate 30 terabytes of data each night of operation for a total of 150 petabytes over the predicted 10-year operational window[14].

The LSST is designed to fulfill several specific purposes. It will help with understanding dark matter and dark energy, monitoring for potentially hazardous asteroids, studying the outer parts of the solar system, tracking transient objects, and studying the formation of the Milky Way[21].

The storage and search functions for the terabytes of daily data will be publicly available. The LSST was a collaborative effort between 34 universities and national labs with funding from the National Science Foundation and the Department of Energy Office of Science[14].

The highest-volume astronomical data project currently in the planning stages is the Square Kilometer Array (SKA)[13].

The SKA is set to be constructed with portions of the array located in South Africa, Australia, and New Zealand. Construction is to begin in 2018 and expected to be completed in 2024.

The data collection would consist of hundreds of radio telescopes and sensors in an array design spread over thousands of kilometers which will ultimately simulate the sensitivity of a square kilometer telescope.

This project would gather 14 exabytes of data each day of operation and store about 1 petabyte of that daily data[8].

Transmission of this data would match the entire data output of the internet in 2013[2].

This project would record radio waves from the 50MHz to 20GHz range in an effort to answer unanswered questions in astrophysics[15].

This data will transform the study of the earliest formations in the universe, pulsar physics, properties and physics of galaxies, the role of magnetism in astrophysics, and astrobiology with data split into non-image and image processing.

Imaging algorithms will need to detect cosmic activity on a scale as short as nanoseconds and record data for events like supernovae and gamma ray bursts.

Non-image processing will perform pulsar detection and experimentation, which requires processing a massive data stream with computational resources searching for variations in the data, which may indicate a possible pulsar[15].

Processing the SKA data will require the use of existing cloud computing resources from nearly all services, and may require rapid advances of desktop cloud computing infrastructure to supplement that of current cloud providers[16].

The SKA is a collaborative effort between ten nations, dozens of universities, and industrial firms.

It will represent the most data-intensive project of any kind thus far[13].

Making use of this data has challenged almost every part of the big data field. The processing of imaging data from astronomics and astrophysics platforms is a process of recording high-definition images of the sky, comparing all parts of this image to preceding and successive images to determine the movement of individual objects, then directing the most likely candidates for real astronomical changes to human experts for classification [11]. The depth and detail of images varies depending on the project goals and wavelength of light being observed. The processing of non-imaging data is a challenge in gathering data streaming from sensors, detecting anomalies in that data, and alerting astrophysicists to potentially important phenomena. Storing and processing astronomic data for searchability has strained the computing resources available to each respective project and prompted developments in computing and storage [13].

3 BIG DATA INFRASTRUCTURE IN ASTROPHYSICS

For ground-based observation projects, the most common source of noise in data which require cleaning are satellites, 'junk' in earth orbit, and defects in the telescope lens which can create artifacts [20].

Two of the most effective methods of cleaning astronomical datasets are the Hough Transform method and the Renewal String Approach [1, 3, 20]. Both of these methods are designed to account for satellites, space junk, or aircraft drawing lines across the sky when observed from the ground in successive images. The Hough Transform is a general data mining technique which searches through data, and maps lines based on the placement of points in the sky. Points in subsequent images are matched against the mapped lines. If the number of points mapped to a line are higher than expected, the points are flagged as possible aircraft or satellites. Renewal String is a method developed specifically for cleaning astronomical datasets. This method searches for line segments formed within a series of images in a given portion of sky imagery. The model compares object movement by renewing images at fixed time intervals and identifying line segments formed by moving objects which identify a noise-generating object[1].

Data curation and storage must be handled in an accessible way. Researchers and space agencies across the globe are contributing to projects, and all must be able to access data and potentially upload portions of data to different cloud and open source storage systems[10, 18]. The mining of astronomical data has created an entirely new field of astroinformatics [4] which focuses on efficient management of computing resources. The resources needed for upcoming astronomy projects, such as the SKA, will require both cutting-edge super computer clusters and cloud computing solutions beyond the current capabilities. Desktop cloud computing networks may provide a complimentary option to help alleviate the computing burden[16]. The computational volume for astrophysics data is now measured in the hundreds of teraflops. The initial processing of data for terrestrial telescope projects is typically performed on site by supercomputer clusters, which format and compact the raw data. Cloud computing methods are applied to user analysis of cleaned datasets after data is stored for user access[16].

Data mining of astrophysical and astronomical data requires search and selection features which can quickly return data relevant to a researcher's needs. Storage and query structures for modern projects, include established tools like MySQL, SciDB, MonetDB, Hadoop, in addition to project-specific query tools that are perpetually in development[19, 26].

Within these query tools there is development of unique algorithms and complimentary features to optimize query results. The ability to analyze astronomical data is ultimately a problem of identifying significant events, new attributes for astronomical objects, and interesting "front-page-news[6]" outliers using query and analytical tools, which must sift through very large and noisy data. There is ample space for research and development in the search and query task with astronomical datasets. The diversity of astrophysics data provides an environment to challenge query algorithms and the scale of data challenges the most robust query methods. There are many examples of experimental tools for the query and analysis process such as MapReduce[11] tools for efficiently performing data reduction across computing nodes; XtreemFS[11], which provides cloud data replication to improve the accessibility of queried data; and Charles[17], a query advisory system which queries the query system itself to provide a user direction on possible research.

4 HOW BIG DATA HAS CHANGED ASTROPHYSICS

Astronomical and astrophysical data is growing rapidly in size, and researchers are able to query increasingly detailed volumes of data as big data tools develop alongside the observational and recording technology.

Many of the leading big data tools in astrophysics and astronomy were developed around the Sloan Digital Sky Survey (SDSS)[8]. SDSS began observations in 1998 and gathered astronomical data as images until 2009. SDSS ultimately generated 140 terabytes of imagery data, which quickly dwarfed the amount of data gathered in the entire history of astronomy[8]. The fields of Astroinformatics and Astrostatistics[9] emerged as data science caught up to this flow of data. SDSS took imaging data of the visible spectrum and ultimately recorded data for a billion celestial objects including stars, galaxies, and quasars. Although an entire decade of SDSS data will be matched every ten days by the LSST, the SDSS was vital at bringing astronomy and astrophysics into the big data era[8]. The machine learning, data processing, storing, and querying methods essential to modern astronomy and astrophysics were largely developed concurrent to SDSS. The big data methods developed alongside SDSS were applied to later projects and even retroactively to the data from earlier sky surveys. The big data applications in astronomy are applied primarily to the terrestrial observation platforms for now.

Orbital platforms are limited by the volume of data that can be broadcast to earth and are difficult to classify as big data projects. They are not, however, beyond the reach of developments in big data in astrophysics and astronomy.

The query, cleaning, and analytic methods common to sky surveys are used with data from space telescopes. The Hubble Space Telescope (HST) generates beautiful astronomical imagery and launched astronomy into the big data era with the Hubble Deep Field[12]. The HST, however, returns only 17.5 gigabytes of data a week and that full data is not made publicly available since it is not considered a big data project. The upcoming James Webb Space Telescope (JWST) will implement data compression and reduction as part of its on-board computing process to format data before transmission[7]. It will also not produce data at a volume or level of openness to be considered as much of a big data project.

5 CONCLUSION

The LSST and SKA are the most prominent research platforms designed around big data tools which emerged from earlier projects. The SKA is so data intensive that some consider it as much a big data project as an astrophysics project[13]. There is little indication that astrophysics and astronomy will become less data-intensive in the future. Instead, these fields will continue to push development of infrastructure at all levels of the data science pipeline. Future collaborative efforts may introduce a true big data paradigm to space telescopes to match the embrace of big data in terrestrial observation platforms.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the I523 TAs for technical assistance and Dr. Geoffrey Fox for the lecture material which inspired this paper's idea.

REFERENCES

[1] et al Amos J. Storkey. 2004. Cleaning sky survey data bases using Hough transform and renewal string approaches. *University of Edinburgh* 347, 1 (Jan. 2004), 36–51.

[2] Ross Andersen. 2012. How Big Data Is Changing Astronomy (Again). (Aug. 2012).

[3] Danko Antolovic. 2008. Review of the Hough Transform Method, With an Implementation of the Fast Hough Variant for Line Detection. *Department of Computer Science, Indiana University, and IBM Corporation* 133 (04 2008), 1539–1548. https://doi.org/10.1541/ieejeiss.133.1539

[4] Kirk Borne. 2009. Scientific Data Mining in Astronomy. *Next Generation of Data Mining (Taylor & Francis: CRC Press)* 0911, 0505 (2009), 91–114.

[5] Kirk Borne. 2014. Top 10 Big Data Challenges fi?! A Serious Look at 10 Big Data Vfis. (2014).

[6] Kirk D. Borne. 2008. A Machine Learning Classification Broker for Petascale Mining of Large-scale Astronomy Sky Survey Databases. *Department of Computational & Data Sciences, George Mason University* 10, 1007 (2008), 5.

[7] Mariel John Borowitz. 2011. The James Webb Space Telescope: A Worthy Investment in Space Science. Space Foundation. (July 2011).

[8] The Economist. 2010. Special Report — Data, data everywhere. *The Economist* The Economist, February 2010 (02 2010), 1–13.

[9] G. Jogesh Babu Eric D. Feigelson. 2012. Big data in astronomy. *The Royal Stastical Society* August 2012 (2012), 22–25.

[10] Megan Gannon. 2014. How Scientists Tackle NASA's Big Data Deluge. (01 2014).

[11] et al Harry Enke. 2012. Handling Big Data in Astronomy and Astrophysics: Rich Structured Queries on Replicated Cloud Data with XtreemFS. *Datenbank-Spektrum* 12, 3 (11 2012), 172–181.

[12] HubbleSite.org. 2016. "Hubble Essentials: Quick Facts". Archived. (July 2016).

[13] IBM. 2012. Square Kilometer Array: Ultimate Big Data Challenge. (2012), 11 pages.

[14] LSST. 2016. Education and Public Outreach (EPO) Completes a Milestone Review, About LSST. (2016). https://www.lsst.org/about

[15] SKA Project. 2017. The Square Kilometre Array (SKA). Jodrell Bank Centre for Astrophysics. (May 2017). http://www.jodrellbank.manchester.ac.uk/research/research-centres/ska-project/

[16] J Tseng R. Newman. 2011. Cloud Computing and the Square Kilometre Array. *Square Kilometer Array* 2011, 124 (2011), 21.

[17] M.L. Sellam, T.; Kersten. 2013. Meet Charles, big data query advisor. *UvA-DARE (Digital Academic Repository)* 6th Biennial Conference on Innovative Data Systems Research, CIDR 2013 (January 2013), 1–8.

[18] Matt Stephens. 2008. Mapping the universe at 30 Terabytes a night: Jeff Kantor, on building and managing a 150 Petabyte database. (10 2008). http://www.theregister.co.uk/2008/10/03/lsst_jeff_kantor/

[19] Matt Stephens. 2010. Petabyte-chomping big sky telescope sucks down baby code: Beyond the MySQL frontier. (11 2010).

[20] Amos J Storkey, Nigel C. Hambly, Christopher K. I. Williams, and Robert G. Mann. 2003. Renewal Strings for Cleaning Astronomical Databases. *ArXiv e-prints* 1408, 1489 (Aug. 2003), 559–566. arXiv:cs.AI/1408.1489 http://adsabs.harvard.edu/abs/2014arXiv1408.1489S

[21] Large Synoptic Survey Telescope. 2010. Large Synoptic Survey Telescope gets Top Ranking, fia Treasure Trove of Discoveryfi. Public Release. (08 2010). RELEASE LSSTC-09.

[22] Tiffany Trader. 2014. Astrophysics: The Icing on the Big Data Cake. Datanami.com. (01 2014).

[23] Australian National University. 2017. Data Release DR1. (06 2017).

[24] Yongheng Zhao Yanxia Zhang. 2015. Astronomy in the Big Data Era. *Data Science Journal* 14 (2015), 11. https://doi.org/10.5334/dsj-2015-011

[25] et al Zachary D. Stephens. 2015. Big Data: Astronomical or Genomical? *PLoS Biology* 13, 7 (July 2015), 2–11. https://doi.org/10.1371/journal.pbio.1002195

[26] Jacob T. VanderPlas Zeljko Ivezic, Andrew J. Connolly. 2014. *Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data.* Princeton University Press, 41 William St, Princeton, NJ 08540.

Big Data Applications in Fraud Detection in Insurance

Mrunal L Chaudhary
Indiana University
Bloomington, Indiana
mchaudh@iu.edu

ABSTRACT

Insurance companies today are incurring a loss in billions of dollars every year because of frauds happening in filing claims, paying premiums, filling applications etc. Detecting frauds manually or by other traditional means is an impossible task since magnanimous amounts of data is getting generated every day and fraudsters change their strategies very quickly. For handling such a humongous amount of data, performing real time analysis on it, and getting accurate outputs; it is imperative that a robust, flexible and scalable technology be used which can detect frauds on the fly. Big data provides just the platform needed to perform analysis of such high complexity.

KEYWORDS

i523, HID205, Insurance, Fraud detection in insurance, Predictive analysis

1 INTRODUCTION

The fact that technology is evolving at the fastest pace in the history of mankind needs no more of a proof than a mere glance of eyes around our surroundings. But like everything else, it is a double-edged knife and this advancement in technology comes at a cost of benefiting the fraudsters of the society. A fraud is a wrongful or criminal deception intended to result in financial or personal gain. And while they are rampant in almost every field, there is no surprise that the field of insurance too has been affected by them. Traditionally, Insurance industry estimated that frauds account for about 10 percent of the total losses incurred by them which is equivalent to billions of dollars, and these numbers are only rising [4]. As these fraudsters are getting smarter, and well equipped with technology, insurance companies are facing difficulties in preventing and detecting the fraudulent activities with the traditional ways. Though the growth of modern technology aids the fraudsters in generating sophisticated fraud techniques, the advancement in technology has enabled better and smarter approaches in detecting fraud. Today when data is getting generated at a break-neck speed and transactions are digitally documented in some format, evidence is just hidden in the data for aiding the investigators to control the fraudulent activities [5]. The question then that needs to be addressed is, how to find that evidence easily [5].

2 BACKGROUND

Most of the insurance frauds are committed in the fields of Health care, auto insurance and workers' compensation. These frauds can either be categorized as hard or soft. Hard frauds occur when someone on purpose fabricates an accident or makes up non-existing claims. Soft fraud occurs when they inflate the amount of a legitimate claim [3]. The parties that are most affected by frauds are the loyal consumers who have to pay higher insurance premiums for compensating the losses from frauds, and medical professionals who are concerned of tarnishing their reputation [4]. The people who do insurance frauds can either be organized criminals who draw large sums of money from fraudulent claims, professionals who add on to legitimate claims or ordinary people who just want to cover the amount of premiums or deductibles [3].

3 REASONS FOR FAILURE OF TRADITIONAL APPROACHES

The natural question then that comes to the mind is, "Why are the traditional approaches to detect fraud inadequate?" The rate at which this voluminous data is getting generated is just impossible to efficiently and accurately process manually. The data sources that get generated were far too large and were changing far too often so as to be helpful in scoring the fraud the traditional way, since they used batch processing which took hours or even days together to run [9]. Moreover, the insurance firms used red-flag method for suspicious claim detection. Thus, using sampling techniques was bound to skip some frauds and introduce errors [7]. Also, the data generated by insurance companies is ever evolving and changing, and the traditional approaches are not sustainable to process data at real time [10]. Thus tweaking of parameters in the fraud detection algorithm was an impossible task since it requires a lot of processing time. Hence, traditional approaches to fraud detection lacked both- the flexibility and the scalability [7]. And lastly, traditional approaches only looked at the structured data since the fraud detection systems were not equipped to handle unstructured data, thereby eliminating a huge subset of data, and making the critical decisions on incomplete information [1].

4 CHALLENGES FACED IN FRAUD DETECTION

It is widely recognized that the volume of data is increasing at breakneck speed. In fact experts believe that the amount of data that has been amassed in the last two years- a zettabyte- is more than the data that has been generated since the dawn of the human civilization [8]. But more astonishing than this is that the volume is only one part of the entire equation. With volume, big data deals with velocity and variety as well. Insurers can access a variety of information that is ever increasing and can be accessed through social media and customer feedback and reports in the form of unstructured data. Also photos and videos are another rich source of visual media information that the insurers can lay their hands over. And this information is ever changing and increasing. Thus, dealing with the velocity and variety of the voluminous data are challenges in themselves [7].

5 ADVANTAGES OF USING BIG DATA IN FRAUD DETECTION

One of the biggest advantage that high-performance analytics allows is the ability to use the ever-changing and increasing data sources previously ignored because of the lack of sophisticated tools for handling big data [7]. With the advent of high-performance analytics, billions of rows of data can be processed in a matter of seconds. Thus, insurers can determine fraud scoring in real time. Insurers can now test certain approaches in real time and constantly tweak the parameters for maximizing the output of the fraud detection algorithms. Since high performance analytics can process a magnanimous amount of data in mere seconds, sampling of data is no more needed. Hence the error introduced by sampling can be completely avoided by using big data technologies [7]. With the help of high performance analytics, advanced models like supervised predictive modeling, data mining, social network analysis, social customer relationship management, etc. can be implemented to improve the process of analysis [10]. With the arrival of big data, the process of fraud detection can be completely revolutionized. 'High-performance analytics showcases ways in which organizations can now capture data and use it to their benefits. It has revolutionized the ways in which companies manage data, especially in fraud [7].

6 ROLE OF ANALYTICS IN FRAUD DETECTION

Traditionally, insurance companies have used Statistical models for the identification of frauds in claiming insurance policies. The problem with the traditional way was that it worked in silos, and hence are not scalable to handle the rapidly growing information from different sources [10]. Analytics therefore play an important role in fraud detection by addressing these issues. The key benefits are

(1) Analytics help in the detection of low key incidence events
(2) Analytics helps in effective integration of data.
(3) Since most of the data related to fraudsters like third party reports, is available in unstructured format, text analysis can play an important role in providing valuable insights in fraud detection [10].

7 INNOVATIVE FRAUD DETECTION METHODS

7.1 SNA (Social Network Analysis)

"SNA allows the company to proactively look through large amounts of data to show relationships via links and nodes" [6]. SNA tool combines many analytical methods such as business rules, statistical methods, pattern analysis, and network linkage analysis for uncovering the data to show these relationships [10]. Take a straight forward case of an accident claims made by a victim. Though it may look simple in the beginning, SNA can reveal that the address given by the victim or the car involved in the accident was in fact used in multiple other claims, suggesting a fraudulent activity [6]. SNA works like this: After feeding the data into an ETL (extract, transform and load) tool, the Analytics team scores the risk of fraud by prioritizing the likelihood based on the history

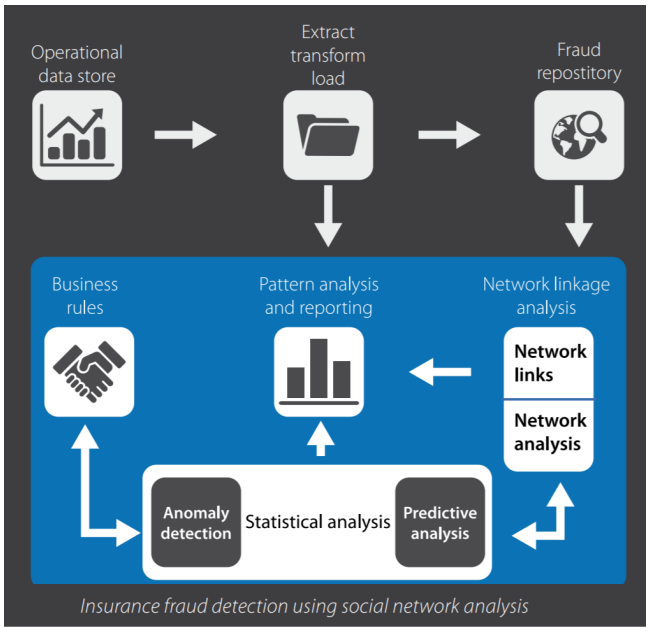

Figure 1: Social Network Analysis Flow chart [10].

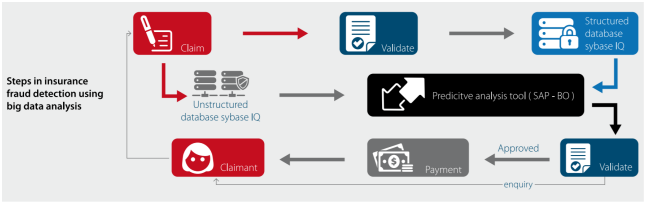

Figure 2: Predictive Analysis Flow chart [10].

of the claimant, relationship with other fraudsters, multiple claims, etc. The Fraud Identification and Predictive Modeling process is an integrated framework of technologies like sentiment analysis, text mining, content categorization, social network analysis. Thus, by doing this, the insurer can rate each claim. A fraudulent claim can therefore be indicated by a high rating. The correctly identified frauds are then added into the business use case system [6]. Figure 1 Social Network Analysis Flow chart .

7.2 Predictive Analysis for Big Data

Predictive analysis makes use of text and sentiment analysis to go through big data for fraud detection [6]. Big data helps in proactively detecting the fraudulent cases by quickly sifting through the large claim reports which are unstructured in nature. An important point to note is that people committing frauds mostly alter their story with time. And these clues are hidden in the log reports submitted by the claim adjusters [10]. The computing system based on the business rules can spot the evidence of the fraudulent claims easily with the help of text and sentiment analysis [6].

Figure 2 Shows Predictive Analysis Flow chart.

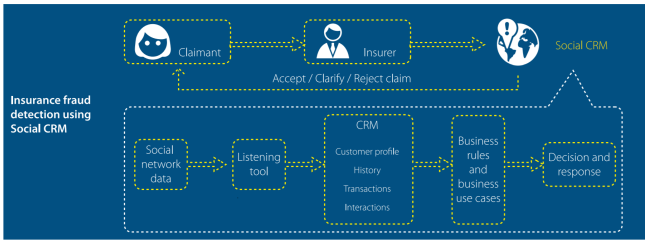

Figure 3: Social Customer Relationship Management Flow chart [10].

7.3 Social Customer Relationship Management

The SCRM is a process that Companies should follow to link their CRM with social media for better understanding of customer behaviour and demands and general trends [6]. The reference data which the company extracts from social media chatter using a 'listening tool', along with information stored in the company's existing CRM is fed into a case management system. This system then sends a response about whether the claim is fraudulent or not, which is confirmed by the investigators [10].

Figure 3 Social Customer Relationship Management Flow chart.

Most of the Insurance Fraud detection tools build a framework around the claim management vertical, but for building a more robust system, a holistic framework is needed, one which can identify potential areas for frauds like in application, premium, claims, etc. Following are the 10 steps for implementing the analytics of fraud detection [10].

8 10-STEP IMPLEMENTATION OF ANALYTICS IN FRAUD DETECTION

8.1 Performing SWOT analysis

Due to the increasing awareness of the fraud detection systems, before going for a solution, the insurance company should perform a SWOT analysis of the existing solutions and choose the most suitable one.

8.2 Building a team dedicated for fraud detection

It is important to form a dedicated team for fraud detection which can handle the responsibilities of fraudulent claims going unnoticed otherwise.

8.3 Building a Solution versus buying it

In case the insurance company decides to go for building a solution, they need to be sure that they have the required skill set for building an in-house analytics product. And if not, then it should find an analytics solution that best fits its requirements.

8.4 Data cleaning

Databases should be integrated and redundancies should be removed from data.

8.5 Coming up with relevant business rules

Certain types of frauds are very industry or company specific and this knowledge can only be gained through experience. The insurance companies should use the existing in-house expertise to define the business rules.

8.6 Defining the thresholds for Anomaly Detection

The insurance companies need to carefully set the threshold for Anomaly detection after considering factors like type of insurance, rate of fraudulent claims, time available etc. If set to a high value there is a chance that many fraudulent claims might go unnoticed and if set to a very low value, it might culminate into wastage of time.

8.7 Using Predictive Modelling

Data mining tools should be utilized for building models that can produce scores for fraud propensity in regard with unidentified metrics. The result thus can then be given for further analysis.

8.8 Using Social Network Analysis

SNA models relationships between various entities involved in claims and therefore, has proved to be very helpful in identifying the fraudulent activities. These entities can be anything from geographic location, car in case of car accidents, age groups, financial status, phone numbers, etc. Through SNA it can be found out that the linkage between some entities is higher than the average connection numbers, thus indicating fraudulent activities.

8.9 Building a Social Customer Relationship Model

The integrated case management systems allows the insurance companies to capture relevant findings to claims data and ensures efficiency and proper assessment of investigations.

8.10 Forward looking Analytic solutions

For building a truly robust system the insurance companies should keep looking for additional third-party sources of data and their integration with existing solutions for increasing the efficiency of the fraud detection system. Also, they should always keep in mind the issue of scalability. With the ever-increasing data, the system should be such that it can handle the size of the data [10].

9 CONCERNS RELATED TO FRAUD DETECTION SYSTEMS

Though Fraud detection system have become pretty robust and sophisticated, there are still come issues that need to be addressed which are the following:

- The data that is collected by the insurers can be properly utilized. The biggest hurdle though in their way is legislative barriers and privacy protection laws that hinder the analysis of the insurance companies [1].
- No matter how advanced the fraud detection tools becomes, there will always be a dependency on humans for converting the reports into actionable intelligence [1].

- Most of the modelling techniques are highly dependent on the past behaviors of the fraudsters. But their behavior changes so quickly that it makes the whole analysis worthless. Evaluating the quality of the data therefore is a huge struggle [2].
- The loses incurred by the insurance firms are somewhat compensated by charging a higher amount for premiums and taking more time, thus the insurance firms may lose out on loyal customers [1].

10 CONCLUSIONS

The upsurge of analytics presents a world of limitless potential for insurance companies which have long held a foundation of information. With the advent of big data, high-performance analytics technology represents an opportunity to completely revolutionize the way fraud is detected. Though Big Data applications in Insurance are still in the early stages, they have proved to be powerful to easily handle and process the velocity with which variety of voluminous data is getting generated. In the years to come, Big Data Analytics has showcases the potential to find widespread applications in the field of insurance.

REFERENCES

[1] Insurance fraud EU. 2016. The Role of Data and Analytics in Insurance Fraud Detection. (June 2016). https://www.insurancenexus.com/fraud/role-data-and-analytics-insurance-fraud-detection

[2] Friss. 2017. The 8 Biggest Fraud Challenges for Insurers. (Feb. 2017). https://www.friss.com/en/news/the-8-biggest-fraud-challenges-for-insurers/

[3] Insurance Information Insitute. 2017. Background on: Insurance fraud. (Sept. 2017). https://www.iii.org/article/background-on-insurance-fraud?cm_mc_uid=96455616751215069824528&cm_mc_sid_50200000=1507445074&cm_mc_sid_52640000=1507445074

[4] Kim Minor. 2013. Improving Claims Fraud Detection in Insurance. (May 2013). http://www.ibmbigdatahub.com/blog/improving-claims-fraud-detection-insurance

[5] Paul Nelson. 2017. Fraud Detection Powered by Big Data - An Insurance Agency's Case Story. (2017). https://www.searchtechnologies.com/blog/fraud-detection-big-data

[6] Sachin Pandhare. 2017. Big data Analytics: new whistleblower on insurance fraud. (2017). https://www.infosys.com/industries/insurance/white-papers/Documents/new-whistleblower-insurance-fraud.pdf

[7] James Ruotolo. 2013. Big Data for Fraud Detection. (May 2013). http://www.insurancetech.com/big-data-for-fraud-detection/a/d-id/1314553?

[8] Jonathan shaw. 2014. Why 'Big Data' Is a Big Deal. (March 2014). http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

[9] Ikanow Editorial Team. 2014. HOW CAN I USE BIG DATA ANALYTICS FOR FRAUD DETECTION? (Feb. 2014). http://www.ikanow.com/how-can-i-use-big-data-analytics-for-fraud-detection/

[10] Ruchi Verma and Sathyan Ramakrishna Mani. 2013. Using Analytics for Insurance Fraud Detection. (Dec. 2013). https://www.the-digital-insurer.com/wp-content/uploads/2013/12/53-insurance-fraud-detection.pdf

Big Data Analytics in Cyber Security and Threat Research

Tousif Ahmed
Indiana University
Bloomington, IN
touahmed@indiana.edu

ABSTRACT

The introduction of big data poses new security threats for the organizations and the consequences can cause catastrophic damages for organizations. Newer threats are sophisticated enough that the existing security mechanisms might be ineffective to thwart the attacks. However, big data analytics and tools can help the organizations to detect the threats and take protective measures. Moreover, big data analytics and machine learning together can detect newer threats which was not possible before. In this paper, we discuss some new cybersecurity threats and challenges that has been bolstered by big data and then we discuss some new big data related security mechanisms which can help the organizations to protect their resources.

KEYWORDS

i523, HID 237, Big Data, Cyber Securty, Threat Intelligence.

1 INTRODUCTION

With the rapid growth of consumer based services, artificial intelligence, and social medias, the number security and privacy threats are also increasing. Emerging cyberattacks are not only a threat for industries, end-users are at more risk than ever. A recent stat published by MalwareBytes suggests that there are a rapid proliferation of number of security threats recently and more than 1 billion malwares were detected in their collected data of six month time span [6]. The consequence of the cyberattacks can be catastrophic to the Organization and the security of the consumers. Therefore, proper defensive mechanism and security controls need to be placed to impede cyber threats. However, newer threats have become so sophisticated that now it's extremely difficult detect and mitigate newer threats. Existing security mechanisms are ineffective due to the large volume of traffic on the internet and lack of resources.

To accommodate and protect from new line cyber attacks, industries now rely on big data analytics. These new big data analytics shows promises for these new sophisticated cyberattacks and a new area of research has established which studies newer threats and explore defensive mechanisms. Advanced analytics and a combination of machine learning can improve the cyber security and provide a new defensive mechanisms. For example, with the large number of traffic on websites big data analytics enables the security administrators to detect intruders effectively and prevent them from compromising the system. Moreover, the big data analytics can also provide newer lines of defense against consumer's security threat. For example, financial industries can analyze the user's financial behavior and detect anomalies which may help users to mitigate the consequence of stolen credit card or identity theft. The introduction big data analytics poses interesting conundrum for security and privacy, on one hand it creates new security and privacy risks and on the other hand big data analytics provides newer tools for mitigating security threats.

In this paper, we first discuss the cyber security and privacy threats of big data analytics on organizations and consumers, then we discuss some potential applications that can be used to mitigate cyber security risks.

2 CYBER SECURITY THREATS AND CHALLENGES

In this section, we briefly discuss various security threats that is posed by the introduction of big data analytics:

2.1 Increased potential for security breaches

Now, organizations are collecting more data which increased the motivation of the attackers to exploit the organization's vulnerabilities and breach their security. The main objective of the attackers are accessing and downloading consumer's data and that data can be sold to other companies or those information can be used to infiltrate more sensitive data [3]. For example, if the user's email and birth date can be revealed from one system, that data can be used to infer other sensitive information like banking information. The availability of the data motivates the attacker to attack a system and gain access. Recent series of security breaches on high profile companies like Yahoo [10] and Equifax [13] provides examples of increased potential, and the number of affected users provides an example of the consequences.

2.2 Threats to consumers privacy

With more data and available tools, consumer's privacy is at more risk than ever. Users share bits of their personal information on various websites and the combined information from various websites poses new security and privacy threats to the consumers. Consumer's privacy threat is a big risk for an organization as the security of the user's data is correlated with the reputation of the organization. Therefore, thwarting the security attacks have become an important issue for the organizations.

2.3 Sophisticated vulnerabilities

With sophisticated big data tools, attackers are now implementing more intelligent spams, malware, and website threats [5].By using machine learning tools, newer generation of spams have been proliferated which is now hard to detect. The rise of chatbots and automated text generating tools have enabled creating spams extremely easy. Moreover, social medias have made spam distribution extremely easy. Besides spams and malwares, increasing number of device usages (smartphone, IoT devices) have increased the number

of BotNets. Using analytics tools, now botnet distribution and management has become pretty easy [4, 11]. The emergence of deep learning has also motivated newer types of security threats.

2.4 Complex security management and monitoring

With high volume of data, now security management is extremely complex. It is very difficult to correctly assess the risks of a system and monitor the networks. With millions of users accessing websites, now it is almost impossible to scrutinize the network traffics. The high volume of data creates additional security risks for organizations. Existing signature based intrusion detection has become irrelevant with higher number of traffics and unpredictable nature of the users.

3 BIG DATA ANALYTICS FOR CYBER DEFENCE AND THREAT RESEARCH

In response to newer security threats, big data analytics have been used to provide newer set of tools to the security administrators. Based on the existing researches and news, in this section we discuss some newer techniques for cyber defence:

3.1 Scalable Anomaly detection

The most widely used big data tools for cyber defense is the anomaly detection. Now, with the help of numerous data it has become extremely easy to detect anomalies. Anomaly or abnormal behaviour detection is pretty easy to detect with large volume of data, as most user's exhibits common behavior or patterns. Illegal or bad actors act differently while accessing a system and using clusters it has become very easy to detect anomalies. Nowadays, anomaly detection systems have been incorporated to detect scammers, credit card theives, hackers, and potential intruders. Network monitoring schemes have become extremely scalable and efficient, so that it can easily raise an alert once an abnormal behavior exhibits [8, 9].

3.2 Effective Malware Analysis

The existing ways to detect malwares are highly inefficient as it highly relies on the previously seen malwares and signatures. Once a software behaves in an inappropriate way (e.g., accessing files that the software does not supposed to, creating multiple copies, logging keys), then the antiviruses generates an alert and then the software is matched with the virus database. With the new attack mechanisms, the malware analysis and reverse engineering of the softwares are highly time consuming and inefficient. Moreover, they do not always help to prevent a security breach. With the help of big data analytics tool, now it's become extremely easy to analyze high volume of software behaviors, network traffics, file-system modification. Therefore, big data analytics shows promises of a more intelligent antiviruses with more effective malware analysis [7].

3.3 Fake user detection and prevention at scale

With the growing number of services, one problem that the organizations regularly face is that the number of fake users. Often fake users create profiles in various platforms and websites. These fake users often create problems on the platforms ecology and exhibits abusive behaviors towards legitimate users. Identifying fake users often extremely difficult with the large number of legitimate users. Now, big data analytics provides various tools to analyze networks effectively which allowed the platforms to detect fake users by analyzing their behaviors. Often these fake user's creates a large networks and by clustering algorithms it has become pretty easy to isolate the group of fake users.

3.4 Spam fighting and detect Botnets at scale

Spam and Botnet are one of the major security problem for organizations. They cost useful resources and the underground economy of spam suggests that spam accounts have high benifits [12]. Everyday hundreds of users falls into phishing attack which cost the users moentary loss. The proliferation of social medias and crowd sourced systems have increased the spam distribution. However, now organizations are extremely effective on detecting spams and botnets. Various data anayltics are helping organizations to prevent spammers and protect naive users [11].

3.5 Automated security management

As mentioned earlier, security management has become extremely complex with higher volumes of traffic and data. However, big data management tools provide better security management and new data analytics and visualization tools provide automated approach of security management. Now, it's not necessary to manually investigate the behaviors and generate rules. Using the tools and machine learning, now it is possible to predict threats and automate the security and risk management.

3.6 Better surveillance and cyber safety

Since 9/11, we have seen an increasing usage of communication technology by terrorists or malicious users. However, with the variety of platforms it is has become hard to identify these malicious actors. Big data analytics tools provide a better tool for government surveillance. Although such massive surveillance compromises public privacy, still such surveillance has become effective to thwart dangerous national attacks and so far more than 50 terror plots have been thwarted [14]. Such massive surveillance have become possible due to the big data tools and analytics. They have been successful detecting anomalous behaviors and identifying the bad actors. Similar to the terrorists, these new tools are helpful for detecting social menaces like pedophiles online and keep people safe online.

4 CONCLUSION

Although big data tools and analytics has created cyber threats, it is also helping defending the threats and shown promises on successful defending in the future. According to recent stats, security breaches are declining with the help of big data analytics [2]. However, with the apparent benefits still companies have not widely adopted big data tools for security, only one in five companies are using big data security at this moment [1]. The main reason for not adopting big data analytics is the high cost and lack of human resources. However, it is expected that the cost will be reduced and more people will be interested on big data related tools which may

influence widespread use of big data analytics. With the increasing usage and better tools, threats will be more sophisticated and defensive mechanisms need to be advanced on parallel.

ACKNOWLEDGMENTS

The authors would like to thank Professor Gregor von Laszewski for helping us with the instruction and resources that was required to complete this paper. We would also to like to thank the associate instructors for being available on the course website all the time and helping us with their answers.

REFERENCES

[1] Bi-survey.com. 2016. Big Data Security Analytics: A Weapon Against Rising Cyber Security Attacks? . https://bi-survey.com/big-data-security-analytics. (2016). Online; accessed Sept 30, 2017.

[2] CSO Online. 2016. How Big Data is Improving Cyber Security. https://www.csoonline.com/article/3139923/security/how-big-data-is-improving-cyber-security.html. (2016). Online; accessed Sept 29, 2017.

[3] Hervais Simo Fhom. 2015. Big Data: Opportunities and Privacy Challenges. *CoRR* abs/1502.00823 (2015). http://arxiv.org/abs/1502.00823

[4] Y. Gahi, M. Guennoun, and H. T. Mouftah. 2016. Big Data Analytics: Security and privacy challenges. In *2016 IEEE Symposium on Computers and Communication (ISCC)*. 952–957. https://doi.org/10.1109/ISCC.2016.7543859

[5] T. Mahmood and U. Afzal. 2013. Security Analytics: Big Data Analytics for cybersecurity: A review of trends, techniques and tools. In *2013 2nd National Conference on Information Assurance (NCIA)*. 129–134. https://doi.org/10.1109/NCIA.2013.6725337

[6] MalwareBytes. 2017. Malwarebytes Releases Global State of Malware Report, Finds 2016 as Year Threat Reality Catches Up to Threat Hype. https://press.malwarebytes.com/2017/01/31/malwarebytes-releases-global-state-of-malware-report-finds-2016-as-year-threat-reality-catches-up-to-threat-hype/. (2017). Online; accessed Sept 27, 2017.

[7] Rahul Dasgupta. 2015. Big data analytics leads the way for next-gen malware protection. http://techspective.net/2015/04/27/big-data-analytics-leads-the-way-for-next-gen-malware-protection/. (2015). Online; accessed Sept 27, 2017.

[8] A. Razaq, H. Tianfield, and P. Barrie. 2016. A Big Data Analytics Based Approach to Anomaly Detection. In *2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT)*. 187–193.

[9] L. Rettig, M. Khayati, P. Cudr-Mauroux, and M. Pirkowski. 2015. Online anomaly detection over Big Data streams. In *2015 IEEE International Conference on Big Data (Big Data)*. 1113–1122. https://doi.org/10.1109/BigData.2015.7363865

[10] Selena Larson.CNN. 2017. Every single Yahoo account was hacked - 3 billion in all. http://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html. (2017). Online; accessed Sept 27, 2017.

[11] Kamaldeep Singh, Sharath Chandra Guntuku, Abhishek Thakur, and Chittaranjan Hota. 2014. Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests. *Information Sciences* 278, Supplement C (2014), 488 – 497. https://doi.org/10.1016/j.ins.2014.03.066

[12] Brett Stone-Gross, Thorsten Holz, Gianluca Stringhini, and Giovanni Vigna. 2011. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-scale Spam Campaigns. In *Proceedings of the 4th USENIX Conference on Large-scale Exploits and Emergent Threats (LEET'11)*. USENIX Association, Berkeley, CA, USA, 4–4. http://dl.acm.org/citation.cfm?id=1972441.1972447

[13] Tara Bernard and Tiffany Hsu and Nicole Perlroth and Ron Lieber. 2017. Equifax Says Cyberattack May Have Affected 143 Million in the U.S. http://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html. (2017). Online; accessed Sept 27, 2017.

[14] The Washington Post. 2013. NSA head: Surveillance helped thwart more than 50 terror plots. https://www.washingtonpost.com/news/post-politics/wp/2013/06/18/nsa-head-surveillance-helped-thwart-more-than-50-terror-attempts/?utm_term=.784e59848c4f. (2013). Online; accessed Sept 29, 2017.

Big Data Analytics in Biometric Identity Management

Robert W. Gasiewicz
Indiana University
711 N. Park Avenue
Bloomington, IN 47408
rgasiewi@iu.edu

ABSTRACT

The United States Government, through its collection and use of biometric data, has leveraged big data in order to protect its citizens and keep our country safe. The speed and accuracy with which this biometric data can be effectively matched to an identity can mean the difference between life and death, as well as the integrity of our institutions. This paper predominantly focuses on how the United States Government, collects, stores, and uses big data to facilitate solving crimes and to enhance national security.

KEYWORDS

i523, HID316, Big Data, Biometrics, Fingerprinting, 2-Print, 10-Print, Matchers, Matching Algorithms, DHS, Homeland Security, Border Security, National Security, Immigration, Terrorism, FBI, AFIS

1 INTRODUCTION

Across the spectrum, big data is rapidly changing the way we do business, the way we live, and the way governments around the world do everything they can to keep us safe in the face of an increasingly dangerous world. Long before the advent of big data, fingerprints were used as a means of forensic identification, but it wasn't until technology had progressed to the point to which these prints could be converted and stored in digital format, organized, and then matched against other stored data and even other databases, that this data truly became useful on the large scale that it is today.

Biometrics technology is changing rapidly, and with it, both the size and scope of data being collected. From 2-print to 10-print, iris to facial recognition, the demand for both data intensive processes and rapid matching have grown exponentially, and understanding how the United States Government uses biometrics is a case study in big data if there ever was one.

2 HISTORY OF FINGERPRINTING: THE ANALOG ERA

In 1858, a man by the name of Sir William James Herschel began using fingerprints as a means of identification [4] near Calcutta, India. This started as a means of not solving crimes, but preventing them; Sir William's aim was to thwart attempts at forging signatures - something that had begun to occur at epidemic proportions. Herschel also used fingerprinting to prevent the collection of pension benefits by relatives after the pensioner had deceased.

It wasn't until 1886 that Scottish surgeon, Dr. Henry Faulds, proposed the concept of using fingerprints to identify criminals to London's Metropolitan Police [3]. Incredibly, they dismissed his proposal.

By 1906, the concept of identifying criminals using fingerprints had made its way to the United States, first in New York City and then elsewhere throughout the country. In 1924, the United States Congress created the Identification Division of the Federal Bureau of Investigation (FBI) and 22 years later, they had processed over 100 million fingerprint cards. By 1971, this number had more than doubled [5].

3 BIOMETRICS ENTERS THE DIGITAL AGE

Before the 1960s and 1970s, fingerprints were stored on cards and expert examiners studied fingerprint features, or minutiae, such as ridges, enclosures, and bifurcations. Fingerprints were then filed according to the Henry classification system [1]. Processing was slow, taking weeks or even months and everything had to be done at one central processing facility. Big Data was perfect solution to this problem.

By the dawn of the 1980s, the completely analog system transitioned toward a more digital platform by storing filing codes on early computer systems. It wasn't until 1986 that the Automated Fingerprint Identification System was released commercially to agencies across the United States Government.

4 AUTOMATED FINGERPRINT IDENTIFICATION SYSTEM (AFIS)

In July of 1999, the AFIS or IAFIS system became a fully automated, nationalized computer system intended for enhanced and rapidly expedited matching capabilities. The AFIS system is not only a criminal and civilian database for fingerprints, photographs, as well as military and civilian data, it is also a matching system, providing either positive or negative identification of prints submitted against its cache of stored records. In addition to biometric identification, AFIS also serves as a means of biographic identification based on pieces of data such as name, date of birth, tatoos, various ID numbers, and other relevant personally identifiable information (PII).

As Simon A. Cole explains in his 2002 book, Suspect Identities: A History of Fingerprinting and Criminal Identification [1], AFIS can work in four of the following ways:

1) 2-print (left and right index finger) and 10-print (all ten of a person's digits) taken from a crime scene, body, or border checkpoint and can be checked against a database of other fingerprints

2) A single latent, or partial trace print can also be checked against a database of other fingerprints

3) A complete 2-print or 10-print image can be checked against other stored latent prints

4) So-called *unsolved* prints, both latent and complete 2-print and 10-print images can be stored in the database and checked against any new subsequent additions.

Today AFIS is the largest biometric database in the world.

5 INITIAL ACHIEVEMENTS OF DIGITIZATION

With AFIS, the original intent of digitizing several hundred million fingerprint cards was to make it easier to do a job that was already being performed manually. As outlined above, it met two requirements: identify fingerprints and serve as a central reporting system on criminal history for the United States Government.

As time went on, AFIS began to earn additional credibility in other areas as well. It not only helped to improve the collection and identification process with regard to latent fingerprints, but it also forced the standardization process by which all fingerprints are collected, stored, and matched against. These standards are known as uniform biometric standards and were essential in enabling various government agencies to share data they collect.

In addition to saving the government and the environment an enormous amount of ink and paper by doing away with fingerprint cards, AFIS has also helped to expedite the pace at which criminals are able to be identified as well as how quickly cases are able to be adjudicated. Lastly, an additional immediately recognized benefit of digitization of fingerprint records has been the rapid improvement of digital image quality needed to more accurately match fingerprints.

6 BIOMETRICS AND BIG DATA

The ever-present question in the world of burgeoning big data is always: *how is this useful?* Often large swaths of data are collected as a part of standard business processes, or, in this case, as a part of criminal investigations and only later are new uses found for the data that's been gathered. As technology evolves new possibilities emerge and stewards of the data find new ways in which it can be used.

There are times, however, in which there are catalysts in addition to the steady march of technological advancement that force us to change the way we look not only our data, but at the world around us. After September 11th, 2001, the United States Congress passed the "Homeland Security Information Act" which with the understanding that information systems for collecting biometric and biographical data were already in existence, must be efficient and should not be duplicated throughout the federal, state, and local governments. The U.S Department of Homeland Security was created in 2002, consolidating many disparate agencies under one roof and one new cabinet level position, reporting directly to the President of the United States.

Subsequent to this, it was incumbent upon the United States Department of Justice (DOJ) to use any means necessary to protect the United States from being subjected to any additional acts of terrorism. To accomplish this the DOJ would need to have other United States Government agencies working together to share information, but foreign law enforcement agencies as well.

7 ENHANCED BIOMETRIC DATA COLLECTION

Biometric Big Data got even bigger in 2003 when the recently formed U.S. Department of Homeland Security created the United States Visitor and Immigrant Status Indicator Technology (US-VISIT) program. In order to meet the ever-increasing demands to preserve and secure our national security, additional measures and enhanced collection at border crossings and at airports was undertaken. Prior to US-VISIT, as had been observed for hundreds of years, paper travel documents and biographical information could be easily forged, various systems were scattered across the U.S. Government and were not well-coordinated, and partner countries did not abide by the same sets of guidelines.

With the creation of the US-VISIT program, the digitization of both biometric and biographic details of individuals coming in and out of the U.S. ensured that these details could not be easily forged or altered. Specifically, the use of fingerprints, and moreover the ability to match them against the largest biometric database in the world in around 10 seconds, prevents untold hundreds of thousands of attempts by dangerous criminals and terrorists from obtaining visas or gaining entrance to the U.S.

By working closely with other agencies across the U.S. Department of Homeland Security, US-VISIT has the same access to crucial fingerprint data as:

1) Immigration and Customs Enforcement (ICE)
2) Customs and Border Protection (CBP)
3) FBI
4) Department of State (DOS)
5) U.S. Citizenship and Immigration Services (USCIS)
6) U.S. Coast Guard (USCG)
7) Department of Justice (DOJ), State, and Local Law Enforcement
8) Department of Defense (DOD) and Intelligence Community

This level of cooperation was solidified even further on October 25, 2005 with U.S. Presidential Executive Order 13388 [2]:

To the maximum extent consistent with applicable law, agencies shall, in the design and use of information systems and in the dissemination of information among agencies:

(a) give the highest priority to

(i) the detection, prevention, disruption, preemption, and mitigation of the effects of terrorist activities against the territory, people, and interests of the United States of America; (ii) the interchange of terrorism information among agencies; (iii) the interchange of terrorism information between agencies and appropriate authorities of state, local, and tribal governments, and between agencies and appropriate private sector entities; and (iv) the protection of the ability of agencies to acquire additional such information; and

(b) protect the freedom, information privacy, and other legal rights of Americans in the conduct of activities implementing subsection (a).

This E.O spelled out the sweeping changes that the U.S. Department of Homeland Security had already made to the way data was collected, processed, standardized, and matched against.

8 THE FUTURE OF BIOMETRICS AND BIG DATA

The future of biometrics and big data is bright. In the past decade, the U.S. Government has moved from 2-print to 10-print, with plans to begin using iris and facial recognition, as well as gait, to identify and neutralize threats. The move from 2-print to 10-print alone represented five fold increase in data storage needs. Storing detailed images of a person's eyes, their face, and the way they walk will require even more data storage capacity and the raw computing power to analyze it. Such advances are necessary to keep us safe in an increasingly dangerous world.

REFERENCES

[1] Simon A. Cole. 2002. *Suspect Identities: A History of Fingerprinting and Criminal Identification.* Academic Trade, Cambridge, Massachusetts. (book).

[2] Information Sharing Environment. 2015. Executive Order 13388. (2015). Retrieved October 4th, 2017 from https://www.ise.gov/resources/document-library/executive-order-13388-further-strengthening-sharing-terrorism-information-protect-americans

[3] Henry Faulds. 1880. *On the skin-furrows of the hand.* Oxford University Press, Chadlington, Oxfordshire, UK. https://doi.org/10.1038/022605a0 (book).

[4] William J. Herschel. 1916. *The Origin of Finger-printing.* Number ISBN 978-1-104-66225-7 in Fundamental Algorithms. Oxford University Press, London. (book).

[5] U.S. Marshals Service Website. 2016. Fingerprint History. (2016). Retrieved October 3rd, 2017 from https://www.usmarshals.gov/usmsforkids/fingerprint_history.htm

Big Data and Artificial Intelligence Solutions for in Home, Community and Territory Security

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

Security is a basic necessity and a major concern of our day to day life. Ideal security systems should effectively and efficiently protect our homes, communities, public infrastructure, cities, and nations from anti social activities. Existing systems and methods haven't reached the level of sophistication to be able to consolidate the large volumes of relevant data from a variety of sources and demographics. The present video surveillance systems use static cameras at fixed locations inside/outside the house to provide alerts when any event detected. However, they are not intelligent enough to understand the context, recognizing the people faces and voices, and differentiate between family members and strangers etc. The limitations of data collection, data mining, and adoption of artificial intelligence led to ineffective systems which are not as predictive as they should be.

The concept of having an intelligent "ear-and-eye" monitoring at home to constantly observe the surroundings both inside and outside can protect the house and people inside house in much safer way. By extending this capability to the neighborhood and city through collaboration would create safe cities across the world. The key differentiating capability from existing systems is to use a micro drone with integrated video and voice with environment sensors to process the voice and facial data with machine learning algorithms. The limited range micro drones can freely move around the house based on the voice and video analytics while learning about family members, friends and strangers.

The technology advancement allows integrating the video, audio and social media data of targeted regions (homes, public places and extended areas) for comprehensive security analysis. Such systems can use advanced statistical methods, image classification and machine learning algorithms to predict and prevent the threats based on the severity probability.

KEYWORDS

i523, HID333, HID337, Artificial Intelligence, Neural Networks, Machine Learning, Micro Drone

1 INTRODUCTION

Information technology and Social network analysis (SNA) are playing significant role in creating safe environments through collecting, processing, analyzing, and utilizing terrorism and crime related data [3]. Intelligence and law enforcement agencies are heavily relying on Information technology and Big data applications in investigating the terrorist activities and criminal networks, suspect subgroups, and their communication patterns.

However, in the present world, security systems are desperately processing data and the decisions/conclusions are being made without considering multiple dimensions of the context. Large corporations, nations, and intelligence agencies are using their individual systems in isolation, but not taking integrated approach to solve the problems in their entirety due to their political and economic interests.

Analyzing individual human behaviors, interactions, transactions, and actions is the key element in identifying the potential threat in advance. Generating and analyzing such data from individual homes and extending the concept to larger groups is the idea behind this discussion.

The current technologies allow to collect data from individual homes and roll up to the communities, cities, and then to the nations across the world. Since this involves with the personal data from people directly, it is required to follow privacy-preservation policies and methods enforced by local/national government agencies. By accessing the household level data of individuals video, voice, social media and other business transaction data would allow to characterize, analyze and assess people's behaviors and motives which can be maintained and processed as needed by Big Data systems. These systems are going to be very complex in nature due to the large volumes, variety and velocity of the data, where the Big Data and Artificial Intelligence (AI) technologies will play a significant role in realizing them. In addition to data collection and mining, if artificial intelligence is applied to analyze and evaluate the data then the crime prediction and prevention would be feasible.

In order to realize such systems, one would need several technologies and sub-systems in various layers to effectively collect, transfer, mining, learning and analyze the data. In the following sections some of the technologies/sub-systems that can be used to achieve the objectives of the proposed conceptual model are described. The discussion here consists of reviewing the available papers/systems related to security informatics and understanding the technologies and methods used. The gaps perceived in the review are attempted to solve by proposing a new concept.

2 HOME SECURITY CONCEPT

This section describes a proposed scalable security system concept, which can be extended to the community, city and beyond. The conceptual model has multiple sub-systems coordinate with each other to establish a robust home security system. In this model, a micro-drone integrated with video and audio will continuously monitor the house both inside and outside. An autonomous dual micro-drone model will have capability to view the surrounding with high resolution frame rates and transfer the data to edge

processing unit and/or cloud based HDFS server. The social media data of house members (e.g., E-mail, Facebook, Twitter, WhatsApp, and other web/mobile applications) gets integrated in to HDFS server.

This will establish a known context with complete information of individuals residing in the house by analyzing the contacts, communication exchange (phone calls, SMS, E-mails), trade transactions, and family/friends/foe information. With the combination of video, voice and social network data a comprehensive home security system can be achieved which not only protects the house but also individuals by having superior knowledge about all the activities. This will require Big Data infrastructure along with the machine learning algorithms in various sub-systems.

This conceptual model can be realized with available technologies and can be architected such that it will become a basic building block for the scalable system.

Some of the existing technology companies making us to believe realization of proposed concept:

- *Squadrone System*: Pioneer in producing intelligent deep learning drones for real-time surveillance
- *Neurala*:A leader in deep learning and neural network software for drones
- *Nvidia*: The world leader in visual computing technologies and leading GPU manufacturer

2.1 Dual Micro-Drones with Video and Audio

The prevailing drone technology is reaching higher levels of sophistication allowing newer concepts to be realized in surveillance applications. In this proposed concept, a micro-drone with integrated video, voice and environmental sensors (temperature, humidity, and accelerometers) can be designed along with learning algorithms to add intelligence. In the basic system, there will be two micro-drones to cover both in-side and out-side of the house (can consider adding more depending on the size of the house/facility) monitoring activities all the time. The drone hardware and software detects and recognize all moving objects through deep learning algorithms such as Regional Convolution Neural Networks (R-CNN). Li Wand and Dennis Sng have reviewed the recent progress of deep learning in object detection, object tracking, face recognition, image classification and scene labeling. Deep models have significantly improved the performance in these areas, often approaching human capabilities. The reasons for this success are two-folded. First, big training data are becoming increasingly available (e.g. data streams from a multitude of sensors) for building up large deep neural networks. Second, new advanced hardware (e.g. GPU) has largely reduced the training time for deep networks [8].

The concept of micro-drone video and audio sub-system is to recognize human face and voice and establish the association. After the human object is created with the face-voice association, human characteristics, behaviors, social contacts, social media accounts, family/friends contact database and personal identification will be mapped. This person object (one of the housemate) will be constantly trained with a large set of data during the learning period. Once the person object is matured with enough intelligence then the system will be ready for monitoring and analyzing the data of the person he/she actually mapped to. Multiple person objects will be created to map all the persons live in that house. The duo micro-drones are intelligent enough to recognize all the persons in the house and understand their behaviors, motives, actions, schedules, plans and their complete activities as time progresses.

These micro-drones freely move around the house to monitor the family, friends, foes, strangers, and people who ever happen to be in the house surroundings and visit to meet housemates. Micro-drones are smart enough to sense the people's emotions based on the expressions, conversations and actions to predict the future consequences and get ready for protective actions (e.g., alerting appropriate people and agencies). Also, micro-drones are equipped with sensors to detect environment conditions like temperatures, wind, rain and humidity etc to take good care of themselves by reaching back to dock/home stations while ensuring that security precautions are addressed.

Since micro-drones are autonomous with self-maneuvering and self-diagnostics capabilities, they will take care of self-charging, protecting themselves from being damaged by staying away from objects and people.

The technologies available to realize such a micro-drone consists of: autonomous multicopters, high resolution built-in 360-degree video cameras, the high speed network link, high speed GPUs, environment sensors, software with the machine learning algorithms for various capabilities discussed above.

2.2 Big Data Infrastructure for Data Handling

In the proposed conceptual model, multiple sub-systems generate the big data from a variety of sources such as video, voice, environment sensors (temperature, humidity, wind etc.). Also big data will be generated from all major social media accounts of individual house mates such as Twitter, Facebook, YouTube, Instagram, E-mails and WhatsApp in addition to GPS location, mobile phone calls and text messages.

The Big Data infrastructure would organize the data through multiple data layers such as collection hub, staging hub and Data Lake. Apache Hadoop has emerged as the de facto standard way of storing all of this *Big Data*, mostly in the form of commercial implementations from HortonWorks, Cloudera and MAPR. Associated technologies such as Flume, HBAse, Hive, Kafka, MapReduce, Spark and Storm offer different ways to get information into and out of Hadoop Distributed File Systems (HDFS) so it can be shared with analytics engines, enterprise applications and user interfaces. Storage should be simple, cheap, recoverable and decentralized to avoid single point of failure.

2.3 Data Privacy Preservation Models

In the proposed conceptual system, multiple layers of sub-systems collect individual home level information for behavioral pattern analysis. This information may include individual person's sensitive personal information. Publishing sensitive information without applying any privacy preserving techniques (de-identification techniques) is going to cause serious privacy issues. The data that will be sent out to be used for next level (community/region) is fed to privacy preservation algorithms like K-anonymization. K-anonymization technique applies generalization and suppression algorithms on data sets so that any single disclosed record

is indistinguishable. One simple example for generalization is replacing phone contact with more generic information like postal code. There are other alternative techniques like t-Closeness and l-diversity can be applied as well to apply different constraints on anonymity. To further improve privacy, fuzzy data techniques like generalize the data, suppress the data and perturb the data can be applied. For social network integration in to the proposed system, models can use subgraph generalization approach to preserve the privacy, which has been discussed in the paper "Privacy-Preserved Social Network Integration and Analysis for Security Informatics".

2.4 Video Data Integration and Analysis

The high quality video image frames will be processed to analyze situational awareness. Learning hierarchical representation of video image data by using deep architecture models is the key component of video analytics. By using the deep learning algorithms to perform image classification, localization, object segmentation, object detection, face recognition and scene labeling would enable to establish a comprehensive situational awareness in the home security context. For example, by using video analytics by recognizing and extracting facial expressions like "happy", "sad", "angry", "scared", "surprised" or "neutral" provide a lot of useful data when it comes to helping intelligence and law enforcement agencies.

This method and approach can be extended to city and region levels by rolling up the data from individual homes. In the context of city and regional security, video analytics would help in people's management, vehicle management, behavior monitoring. For example, in public places video surveillance with deep learning enabled systems can perform constant monitoring, face detection, crowd detection, motion detection, vandalism detection, queue management, people counting, vehicle classification, traffic monitoring, license plate recognition, road data gathering etc. With the advent of new technologies in computing speed there are several Graphics Processing Units (GPU) integrated with high quality image sensors introduced by technology companies such as NVidia can be used in the conceptual model.

2.5 Voice Data Integration and Analysis

The live voice recording integrated with video analysis provides better and accurate insight in to situation awareness for predicting and preventing the potential threats much faster. Traditional voice analytics tools rely on keywords and phonetics. These solutions are not well enough in deriving context and relevancy. With big data and AI advancements, now it is even possible to analyze for things like stress levels, lies, emotional content and more from audio data. Gaussian Mixture Model (GMM) is one of the well known technique for voice recognition. However latest AI and deep learning methods are more accurate classifiers for speech recognition and they are slowly replacing Gaussian mixture Model for speech recognition and feature coding. Google's Speech Recognition API built using deep learning neural network algorithms is the one of the voice analytics software available in the market, which can be used in the proposed conceptual model.

In the proposed conceptual model, the complete characterization of housemates can be performed using deep learning algorithms. This will help to recognize the voice of the persons within the house and build the context. Also, the learning algorithms continue refining the voice characterization of the persons and extend the voice database to other family members and friends. This key aspect of associating voice to the person would help resolving the contextual issues if any arises during behavior assessment.

2.6 Social Media Data Integration

In the conceptual model, along with the video and voice association, if the individual social media activity is monitored his/her behavior can be predicted to assess the motivations and potential actions. The social media accounts can be integrated with the big data system to collect data from applications such as Facebook, Twitter, Instagram, WhatsApp, E-mails, and SMS etc.

By observing and analyzing behaviors on social media, these behaviors can be categorized into the individual and collective behavior. The behavioral data generated by user activities in social media interactions helps in finding personality types and predicting individual behavior. It provides a new dimensional data and by running social media analytics tools like *IBM Watson Analytics* one can predict individual and collective behaviors of people.

Natural language processing (NLP) algorithms along with reasonable quantity of training data can lead to understand sentimental behavior, which is one of the key elements for security informatics. This capability can be applied to the proposed conceptual model to ensure that system is analyzing the social network data.

2.7 Learning Algorithms and Predictive Analysis

The two critical machine learning algorithms needed to realize the proposed concept are one for the face recognition and another for the voice recognition. Deep learning models are potential candidates for these two tasks. Deep learning architectures have four major variations

- Deep Belief Networks (DBN) [2]
- Convolutional Neural Networks (CNN) [4]
- Deep Boltzmann Machines (DBM) [5]
- Stacked Denoising Auto-Encoders (SDAE) [7]

But the best fit model for the proposed system is Convolutional Neural Networks (CNN). CNN convolves learned features from input data, and uses 2D convolutional layers, making this model best fit for processing images.

Predictive analytics is a data mining solution consisting of statistics, machine learning algorithms to determine patterns and future events. Predictive analytics may not be able to tell exactly what will happen in the future but it forecasts what could happen in the future with an acceptable level of reliability [6]. The ability to predict the occurrence of crime events before even it occurs is the key feature of the proposed system. Criminal's behaviour and criminal events frequently can be categorized and modeled. With the the help of predictive analysis technology, it is possible to estimate or forecast occurrence of future events. Therefore predictive analytics are finding use in home security and public safety applications. Significant advances have been made in the integration of predictive modeling with social and behavioral factors, in both equation-based approaches, and probabilistic evidentiary reasoning approaches [6].

3 COMMUNITY DRONE NETWORK

The intelligent drone home security system would enable to provide comprehensive situational awareness at home level. The proposed drones are limited in their coverage area which is strictly enforced by regulatory/intelligence/government agencies. Since this inteli-drone is scalable to extend the coverage by just adding another device, it can be conceivable to create a network of inteli-drones to cover a given community. The community drone network is the collection of security drones covering a specific region within a city which will ensure that the relevant data is delivered to law enforcement and intelligence agencies. This would require one of the drones in the network to be nominated as *Gateway Drone* to communicate with law enforcement/intelligence agencies. Each drone will have the capability to become a *Gateway Drone* as needed. When the new drone is installed it will automatically look for the existing *Gateway Drone* in that community, which if exists then it will join the network and gets registered. If no *Gateway Drone* is recognized, the new drone claims or becomes *Gateway Drone*.

3.1 Gateway Drone

The *Gateway Drone* represents a specific community, which will maintain all the home addresses within that community along with associated personnel as per privacy preservation policies set forth by the regulatory/intelligence agencies. The *Gateway Drone* performs dual function (1) ensure that constantly communicates with *Police Drone* or *City Drone* and (2) monitor its own house security aspects.

The *Gateway Drone* is the critical drone in the regional/city security context as it will provide all sensitive information timely to alert the agencies with the potential threat.

The *Gateway Drone* will discharge or transfer its role when it is no longer capable of doing so due to any technical and/or any other issues. When the existing *Gateway Drone* is dropped off from its role then all the drones within the network will be alerted and one of the drones that is closer to the *Police Drone* or *City Drone* will become the *Gateway Drone*.

4 CITY/EXTENDED REGIONAL DRONE NETWORK

The proposed conceptual model defines the city level security network as a combination of multiple *Community Drone Networks* together. In a given city there can be 'n' number of *Community Drone Networks* based on households, public places, and commercial entities. A network of *Gateway Drones* forms as a *City Drone Network* with one of the drones nominated as *City Gateway Drone*.

Developing a fully autonomous and cooperative multi-drone system requires robust inter-drone communication [1]. There has not been enough research to say with conviction what design would work best [1]. The reliability and bandwidth requirements from the drone networks are diverse. The drone networks, therefore, have all the requirements of the mobile wireless networks and more. Node mobility, network partitioning, intermittent links, limited resources and varying QoS requirements make routing in drone a challenging research task [1].

In the proposed conceptual model, since each drone will use WLAN infrastructure mode in addition to Adhoc mode, there will always be a reliable network available to exchange information. The security drones will switch between Adhoc and infrastructure modes based on the network availability to pass on the information to *Gateway Drones*.

4.1 Drone Networking Challenges

The main challenges that drone networks facing are routing, seamless handover and energy efficiency. Routing has unique requirements - finding the most efficient route, allowing the network to scale, controlling latency, ensuring reliability, taking care of mobility and ensuring the required quality of service. In drone networks, additional requirements of dynamic topology (with node mobility in 2-D and 3-D), frequent node addition and removal, robustness to intermittent links, bandwidth and energy constraints make the design of suitable protocol one of the most challenging tasks [1].

The handover latency and the packet loss during handover process may cause serious degradation of system performance and QoS perceived by users. IEEE has standardized Media Independent Handover (MIH) services through their standard IEEE 802.21. These services can be used for handovers and interoperability between IEEE-802 and non-IEEE-802 networks, e.g., cellular, 3GPP, 4G. MIH, however, does not provide intra-technology handover, handover policies, security and enhancements to link layer technologies. However, MIH is a nascent technology that has not been widely deployed and evaluated [1].

In drone networks, management of energy consumption is an important task. Reducing the energy consumption helps to increase network lifetime, drone payload and flight time [1].

5 CONCLUSION

In this discussion it has been perceived that existing security informatics systems are desperately implemented and consolidation of data and analysis at various layers hasn't been done efficiently. Considering that big data technologies are robust enough to collect the large volumes of data from the variety of sources, a conceptual model is proposed to discuss the feasibility of integrated video, voice, and social media data of individuals to be collected and analyzed for applying the machine learning algorithms. With the technologies such as high speed computing and big data infrastructure, learning algorithms can be applied to solve face and voice recognition. The combination of video, voice, and social network data the proposed conceptual system can address some of the prevailing home, community and territory security challenges and issues.

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] Lav Gupta, Raj Jain, and Gabor Vaszkun. 2015. Survey of Important Issues in UAV Communication Networks. *CoRR* 18 (11 2015), 1–1. https://arxiv.org/ftp/arxiv/papers/1603/1603.08462.pdf

[2] G. E. Hinton. 2009. Deep Belief Networks. *Scholarpedia* 4, 5 (2009), 5947. http://www.scholarpedia.org/article/Deep_belief_networks

[3] Paul Kantor, Gheorghe Muresan, Fred Roberts, Daniel Zeng, Frei-Yue Wang, Hsinchun Chen, and Ralph Merkle. 2005. *Intelligence and Security Informatics* (1 ed.). Vol. 3495. Springer-Verlag Berlin Heidelberg, Atlanta, GA, USA.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Lake Tahoe, Nevada, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[5] Ruslan Salakhutdinov and Hugo Larochelle. 2010. Efficient Learning of Deep Boltzmann Machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Yee Whye Teh and Mike Titterington (Eds.), Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy, 693–700. http://proceedings.mlr.press/v9/salakhutdinov10a.html

[6] Antonio Sanfilippo, Nigel Gilbert, and Mark Greaves. 2012. Technosocial predictive analytics for security informatics. *Security Informatics* 1, 1 (22 Aug 2012), 8. https://doi.org/10.1186/2190-8532-1-8

[7] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* 11 (Dec. 2010), 3371–3408. http://dl.acm.org/citation.cfm?id=1756006.1953039

[8] Li Wang and Dennis Sng. 2015. Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey. *CoRR* abs/1512.03131 (2015), 8. https://arxiv.org/pdf/1512.03131.pdf

Big Data in Sports Visualization

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

ABSTRACT

We discuss Big Data Analytics in "other sports" (not baseball) and how it is being used to improve and evaluate performance via visualization of sports data

KEYWORDS

sports, analytics, data visualization, spatial rendering, HID105, I523

1 INTRODUCTION

Data visualization in sports other than baseball is a rapidly growing field due to the explosion in amount of data captured about these sports. A general problem is that it is very difficult to quantify most team sports because of the chaotic nature of the placement of the ball and of players. Furthermore, it is difficult to make use of massive amounts of data for real sports analysis. Big data allows vast quantities of data points to be assessed in a small amount of time. One example of a use case is SoccerStories, a data visualization tool for soccer which makes use of several different techniques to represent real world events in a series of diagrams. Snapshot is a similar example that uses complex formulas to turn huge quantities of data points into easily readable representations using graphs, symbols, and heat maps. Directorfis Cut is another program that enhances simple statistics with more useful context. There are various methods and tools for sports data visualization.

2 SHOT CHART VISUALIZATION

Team sport data analysts are met with the challenge of the dynamic, and even chaotic position of players and the ball or puck. Data visualization, however, allows decisions to be made based on large data sets that would otherwise be difficult to understand. Snap Shot is an example of a hockey data visualization technique that allows teams to identify the position, trajectory, and effectiveness of shots throughout the course of a game or season [6]. A specific use case was to analyze several theories about "sweet spots" on the ice, or positions that a shooter is more likely to score from. One such theory is that goalies tend to be right handed, and that makes it more difficult for them to block shots from their right side due to holding their stick in that hand [6]. This is an intuition held by many high level coaches that was proved to be false after numerous queries through SnapShot [6]. Similarly, NBA shooting was modeled by experts at MIT's sloan sports analytics conference [3]. Using color and size to represent different data points for shots taken layered onto different positions on the court, one can quickly see where players tend to shoot from and where they are most effective. This type of analysis could be done to show coaches and players where they should be shooting from or what types of plays to draw up for which players.

2.1 Basketball Visualization

Shooting is not the only use of spatial relationships in basketball. By associating spatial data with statistics such as fouls, defensive effectiveness, or rebounds, other types insights can be made. For example, if an opposing player is in foul trouble, a coach can draw up a play to isolate that player in the position where they are most likely to foul again through use of a visualization tool. Furthermore, by layering different spatial visualizations on top of one another, more insights can be made. For example, visually displaying an individual player's shot chart along with an opposing team's defense chart allows predictions to be made on which players will succeed or not succeed in scoring against a given team[3]. Although CourtVision can not effectively do this right now, it contests that this type of chart combination allows users to create and communicate insights into performance more effectively: "almost anyone can understand a well-designed map or chart."[3]

3 TENNIS

Tennis analysis is also being transformed by the massive influx of data into sports, and all this data is represented visually in a variety of ways. Serves can be represented by lines to identify patterns in placement and height from an opponent's serve. Moreover, by associating serves with whether the point was won or lost can identify strengths or weaknesses from areas in service. Andy Murray, for example, used a tool called Hawk-Eye to analyze a loss in a tournament finals against his opponent, Kei Nikishori. Using this tool, he realized that almost every time he left a serve short he lost the point. As Game Set Map observes, "if you're serving short 2nd serves to Nikishori at important points than Nikishori is going to be all over the return and you'll be playing catch up all game!"[1] Using this observation, Murray pushed his 2nd serve farther back in his next matchup with Nikishori and ultimately winning the match. While these types of changes may seem insignificant, players and coaches alike place high value on these sorts of insights. Andy Murray is just one example of a frequent user of Hawk Eye. In fact, Murray took it a step too far when he was caught filming a practice session against an opponent without his permission.[1] Andy Murray clearly places great value on the data and visualization techniques used in tools such as Hawk Eye.

4 SOCCER

Similar to hockey or basketball, soccer is a difficult sport to quantify due to the seemingly subjective methods of evaluating different plays and the randomness of the locations at which plays start and end. One of the most basic and commonly employed data visualization techniques for soccer games is a timeline [5]. This method takes advantage of a predetermined statistic in the game:

the length of each half. By using symbols to represent real world events, such as a ball representing a goal, and placing these symbols along the timeline, one is able to gain a limited understanding of the events in the game. This might give simple statistics such as shots, possession, or fouls, more context within the flow of the game and allow a user to gain more insight into what is happening.

4.1 Field Position Identification

Deeper analysis of a soccer game typically relates a game event such as a shot or pass with positions on the field. According to soccerstories, "the soccer field is the primary object of observation and analysis in soccer. Analysts construct their mental model over the spatial arrangement of the team, and its motion, over time."[5] This type of data can be obtained through the use of wearable technology or through video. One common method of this visualizing this data is a "heatmap, through which player's most frequent positions is displayed by density" [5]. A heat map allows for an intuitive and instantaneous evaluation of a playerfis positioning which might otherwise take the length of a match to evaluate. This technique could also be used to identify a player that does not run back to play defense or a player that gets pulled out of position easily. By identifying visual patterns to make insights such as these, teams can gain competitive advantages [5].

4.2 Set Piece Analysis

An important part of the scouting report for a soccer game is corner kicks and free kicks. According to researchgate, about 30 percent of goals in soccer come from these situations [2]. Identifying patterns in where a team likes to direct set pieces allows a coach to set up their team in an advantageous position. For example, by using a heat map to show the frequency and effectiveness of crosses to different positions from these situations allows coaches to put a zonal marker in position to neutralize the threat. Furthermore, this method can cut back on time spent analyzing video or emphasize insights gained from watching video. By identifying which free kicks or corner kicks were passed to a player very close to the initial position of the ball, a coach or player can quickly prepare themselves for any "trick plays" an opponent has used in the past.

4.3 Flow Graph Uses

Another method of relating simple statistics to locations on the field is a flow graph, "where the size of the nodes shows player's role in the game and the links show the connections between players" [5]. A flow graph relates simple statistics about individual players to other individual players and complex insights to be made quickly, such as discerning which players like to pass to one another or which player has a greater impact on the game [5]. This could also be useful in evaluating a teamfis tendencies such as identifying an inability to attack down the right side or give up more shots on the left. These types of tendencies can be used to make decisions such as what formation to play or what spaces to run into on counterattacks.

5 TEAM SHAPE ANALYSIS

There are several ways team based analysis can often be utilized in data visualization. Director's Cut creates an analysis of a team's "back four," or defensive line, which can be portrayed by simply drawing a line connecting each of the defenders. A team's coach can use this line to identify situations where the back line maintains good flat shape defensively in order to play an offside trap, or alternately situations where the defense gets stretched and could allow a player to get behind the defense [4].

5.1 Player to Player Spatial Relationships

A playerfis proximity to opposing players is another factor that could be useful in evaluating performance. Directorfis cut, for example, breaks down player proximity into three separate categories: no pressure, weak pressure, and strong pressure. This is done by segmenting the soccer field into one meter by one meter squares. Each square is then assigned several attributes regarding the closest player and their speed, direction, and proximity to the square. "Pressure" on each player can then be determined by viewing the attributes assigned to the square he or she occupies [4]. This can be extremely helpful in analysis of an individual playerfis ability to cope with situations that tend to force mistakes. For example, a playerfis pass completion percentage, the number of passes they complete divided by the number of passes they attempt, can be a useful but slightly misleading statistic. A forward will tend to have lower pass completion percentage than a defender, for example, due to the closer proximity of the other teamfis defenders. By associating pass completion with no, weak, or strong pressure, however, individual players can be analyzed and compared to one another more easily. Furthermore, a player that misses more passes under no pressure can be shown what they are doing wrong and correct the problem very quickly. By analyzing what choices lead to passes that end up being cheap giveaways, a player could quickly improve on an important aspect of their game. A player missing passes under no pressure is often likely to be making bad decisions more than lacking talent or technique. By displaying what areas of the field and what distances that poor passes are made over, a player can be better informed and make better decisions on the field. Visualization, in this case, could rapidly accellerate the transition from data to decisions.

6 CONCLUSIONS

In conclusion, data visualization is a widely used application of big data in sports. Technology such as player trackers and video analysis allows players, coaches, and managers to gather and communicate insights into sports performances. Many of these visualization techniques are combinations of simple statistics and context such as field position or time remaining. Basketball and hockey use data visualization for analysis of high percentage shots, while soccer focuses on tactics, formations, and player evaluations. There are numerous tools for each of these, all with various methods of carrying out these tasks.

7 ACKNOWLEDGMENTS

I would like to thank Gregor von Laszewski for generously fixing compile errors and Juliette Zerick for providing accurate and specific feedback.

REFERENCES

[1] 2016. Game Set Map. (2016). Retrieved Oct 23, 2017 from http://gamesetmap.com

[2] Ali Onur Cerrah. 2016. Quantitative Analysis of Goals Scored from Set Pieces: Turkey Super League Application. (2016). Retrieved Oct 7, 2017 from https://www.researchgate.net/publication/307553842_Quantitative_Analysis_of_Goals_Scored_from_Set_Pieces_Turkey_Super_League_Application

[3] Kirk Goldsberry. 2012. CourtVision: New Visual and Spatial Analytics for the NBA. (2012). Retrieved Oct 8, 2017 from http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf

[4] Thorsten Breitkreutz Manuel Stein, Halldr Janetzko. 2016. Director's Cut: Analysis and Annotation of Soccer Matches. (2016). Retrieved Oct 8, 2017 from http://ieeexplore.ieee.org.proxyiub.uits.iu.edu/document/7579433/#full-text-section

[5] Charles Perin. 2013. SoccerStories: A Kick-off for Visual Soccer Analysis. (2013). Retrieved Oct 8, 2017 from http://ieeexplore.ieee.org.proxyiub.uits.iu.edu/document/6634087/#full-text-section

[6] Hannah Pileggi. 2012. SnapShot: Visualization to Propel Ice Hockey Analytics. (2012). Retrieved Oct 7, 2017 from http://ieeexplore.ieee.org.proxyiub.uits.iu.edu/document/6327288/

Big Data Analytics in Sports - Track and Field

Mathew Millard
Indiana University Bloomington
Bloomington, Indiana, USA
mdmillar@indiana.edu

ABSTRACT

Over the years, sports analytics has become more prominent in the way sports are evolving. Data is the driving force in how these improvements and changes are being made. With the presence of Big data and the impact it has on how sports decisions are being made, the importance on how this data can be used and how it can be presented in a useful manner have become a focal point for furthering athletics. This focus on data can be seen all over the professional sports world with many teams hiring data scientists and analysts to use data in order to get the most out of how they develop their rosters. Big data analytics within sports can cover a broad spectrum of topics, but the focus here is to dive deeper into the sport of track and field. First, we will take a look at what track and field looked like without the use of big data. Then, we will dive into the impacts big data has had on the development of track and field.

KEYWORDS

i523, hid216, Track and Field, Big Data, Sports, Running

1 INTRODUCTION

When people think about big data analytics in sports, many think about sports such as basketball, baseball, and football because of the popularity and the volume of statistics used in those sports. Track and field, however, can benefit greatly from the same treatment that other sports have received. Although the sport of track and field is not a widely popular sport, there are many who take part in the sport due to the accessibility of being involved. Events on the running side of this sport range from the one hundred meter dash all the way up to the longer distance races such as the ten thousand meter run with many different types of races in between. On the field event side, there are various events including long jump, high jump, javelin, and many more. With all of these different events, there is a strong sense of specialization among athletes. An attention to detail and technique in the many facets of the sport is what drives the need for data analytics and statistical analysis.

2 TRACK AND FIELD BEFORE BIG DATA

Like all sports, track and field had to start somewhere and it was much different before the world of technology and analytics gave way for a much needed boost. In the days before technology became prevalent, statistical analytics were harder to come by and much of what was being done in the sport was based on theory and vague understanding of how the body worked. One of the most notable aspects of the sport that has come a long way since the injection of data analysis is the footwear that the athletes train and compete in. Back before Nike, Adidas, Brooks, and other companies became prevalent in the track and field scene, the footwear that the athletes used were more simple with less focus on how the shoe can help the athlete perform. This approach ultimately leads to many athletes being prone to injury and we can observe that today in current conditions when athletes get injured from wearing the wrong type of shoes for a prolonged period of time. Unfortunately, the data and analysis just wasn't prevalent at the origins of the sport which could be a reason for why times and marks in events improved drastically as big data and technology improved over time. Another area that benefited from the introduction of big data is overall understanding of fitness. There were plenty of superstitions and theories when it came to training for track athletes, but the aggregation of big data wasn't there in order to take into consideration the intensity and volume a track athlete's training should be at to push their body to the limit. In addition to these areas, another disadvantage in the era before big data alongside the lack of technology in general was that it was a difficult and cumbersome process to get meet results and use that in a productive way. Before big data, ways to gather masses of results to compare and rank individuals quickly and easily just did not exist. This clearly made analysis and prediction much more difficult than it is now and most likely caused a fair amount of confusion, especially at a lower, more unstructured level such as high school track and field. The list of inefficiencies could go on and on when talking about the life in track and field before big data analytics. There are certainly still some imperfections in the sport today, but the impact of big data has numerous positive effects.

3 THE IMPACT OF BIG DATA ON TRACK AND FIELD

As big data started to have an impact on the entire sports world, track and field saw many advancements in various forms such as shoe advancement, understanding fitness, form analysis, and result aggregation. With all of these aspects combined into one package, the sport of track and field not only saw improvements in times and marks in running and field events, but we can see where big data has benefited with the health and injury prevention of athletes as well.

3.1 Shoe Advancement

One of the most interesting and complex portions of current day track and field, especially in distance running, is the shoe development and the engineering that goes into the production between many brands and models. Whether it be Nike, Brooks, Adidas, Asics, or any of the other shoe companies with a serious stake in the running shoe market, the main goal is to provide the athlete with a shoe that enhances performance in a comfortable and efficient manner. This has led each company to come up with various technologies of their own using big data and analysis over time. Much of what is done and the many failures we will never see and this is perfectly

modeled by what Nike recently achieved in their attempt to prepare athletes such as Eliud Kipchoge, Zersenay Tadese, and Lelisa Desisa to break the two hour barrier in the marathon by creating special shoes and modifying training based on data collected and analyzed. In Nike's own words on their website, "during what was called Camp One, we brought these three Breaking2 runners to Nike for extensive testing, gathering data to guide the development of their respective shoes" [1]. Here, we get a brief look behind the closed doors at Nike's special facility for shoe testing. After collecting data and running tests, they successfully created one of the most controversial and fascinating running shoes in years. They call it the Nike Zoom Vaporfly Elite and is only given to elite athletes sponsored by Nike, but later released a few modified versions to sell to the masses based on more testing. Of course, the data collected was based on three runners in this case, but these companies make many more shoes adhere to the common man and woman's needs based on much more data and testing. This, however, shows the willingness of a company to put in the time in order to put big data analysis to the use in order to make shoes that propel the sport forward to new heights.

3.2 Understanding Fitness and Training

Aside from making sure that track and field athletes are training in the right shoes, priming the body with the correct training is another area in which big data is making progress much easier. If you can't prepare your body to go the distance, jump farther or higher, and throw further then fancy shoes with a lot of tech are not of much use. Fortunately, with the rise of smart watches in athletics, there is an abundance of biometric data being collected on a constant basis. Many other procedures such as the VO2 max test and more are used to collect data based on any given person's capacity for endurance activities, but the use of technology such as a smart watch can give us instant and remote access into an athlete's biometric data over many exercises which can clearly tell a much bigger picture. In August of 2017, Business Insider published an article titled "Here's how people are using their smartwatches" which gave some insight into what most people use their smart watches for. Although the main usage was for notifications and text, they found that activity tracking is second leading function that users utilize [2]. Although these results do not cover track and field athletes specifically, athletes are much more likely to use these functions if a coach requires it. This aggregation of data collected by a wearable device like a smart watch allows for coaches, physical therapists, and others to understand how a track and field athlete's body responds under different circumstances. Big data analysis being used in this way allows for better training plans based on athlete fitness which leads to healthier and safer training overall.

3.3 Form Analysis

On top of the rise of wearable devices, technological advances in track and field also came in the form of better video capture that provided more big data to analyze. Form analysis and improvement is another aspect of track and field that has benefited greatly from big data analytics. Recording video of an athlete's performance in various events from running to throwing is important in analyzing form to find areas that have room for improvement in technique. A video analysis and data tracking tool called Hudl has made big changes in how we approach film in sports in general. Track and Field coaches all over from the high school to college level and beyond are utilizing this tool for many things including form analysis. In 2016, Fast Company published an article titled "How Hudlfis Mobile-Video Software Is Transforming Sports". They assert that technology such as "wearable technologies, high-speed cameras, Doppler radar, and data-collection devices" exist to allow for the measurement of many complex movements and techniques [3]. Coaches across all sports and disciplines already take advantage of this technology and it has been a driving force in assisting track and field coaches give proper feedback on form and technique correction that their athletes can make for better results. In some events such as long jump, high jump, and pole vault, tweaking form can lead to drastic improvement in a short time.

3.4 Meet Results Aggregation and Athlete/Team Rankings

Another current day impact that big data has had on the sport of track and field is the improved aggregation of results in a useful and meaningful manner. With track meets and competitions happening every day in locations all over the world, having a place collect all of these results and sort them out is a game changer. It allows for quick analysis and comparison of performances that help rank the best athletes and teams among a nation. This is exactly what tfrrs.org tackles and handles well for United States college track and field and cross country across all divisions. The website collects results from all over the nation and displays it in a well organized manner while using the individual and team results to calculate rankings of the best individuals and teams in the nation. Progress like this is what drives competition and improvement. It allows for athletes, coaches, and fans to make simple comparisons to have more awareness of competition at a larger scale.

4 CONCLUSIONS

Looking back on the major impacts that big data has had on a sport most people gloss over such as track and field, it is difficult to imagine sports living on without big data. In many ways, track and field has made exponential growth in recent years thanks to big data analytics. We often take for granted things like the shoes on the shelves, access to useful smart watches, and websites that simplify our lives, but taking a deeper look should bring about some appreciation for what big data can be responsible for. Sometimes, we see big data as no more than a buzz word to gravitate attention towards something new and fascinating, but here we see the power it has to make a difference. Although, there is always room for improvement in any field and that is still the case here. A lot of the advancements we discussed were fairly recent and could most certainly lead to bigger and better conventions down the road. In the world of track and field, everyone is always looking for more tools and strategies to get faster and stronger. Not only have we made a case for strong involvement of big data in the growth of track and field, but we can make a case that big data should have a bigger role in every aspect of the sport going forward.

ACKNOWLEDGMENTS

The author would like to thank Professor Gregor von Laszewski for providing the opportunity to explore a topic of deep interest.

REFERENCES

[1] 2017. Breaking2 Behind The Innovation. Website. (2017). https://www.nike.com/us/en_us/c/running/breaking2/breaking2-product-innovation

[2] Caroline Cakebread. 2017. Here's how people are using their smartwatches. *Business Insider* (AUG 2017). http://www.businessinsider.com/most-used-smartwatch-features-chart-2017-8

[3] Matthew Shaer. 2016. How Hudlfis Mobile-Video Software Is Transforming Sports. *Fast Company* (FEB 2016). https://www.fastcompany.com/3056061/how-hudls-mobile-video-software-is-transforming-sports

Big Data Analytics in Sports - Soccer

Rahul Velayutham
Indiana University Bloomington
2661 E 7th Street Apt H
Bloomington, Indiana 47408
rahul.vela@gmail.com

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. The impact of Big data in the field of sports, in particular, soccer and how they have helped football clubs evolve their business models and operations from a more hands-on approach to applying complex software and ML models to improve tactics, scouting, and training practices. This study takes a look at the technologies that have been used like MiCoach, Tracab and a look at the leading players like Opta and how the data generated from these companies could be put to use. It is hoped that this study will help demonstrate the importance of Big data in sports, its applications, and avenues for improvement in the field.

KEYWORDS

Big Data, Soccer , Scouting, HID232 ,I523

1 INTRODUCTION

Big Data has become a crucial part of soccer. Data obtained from big data technologies is used to chart training sessions, shape tactics, predict odds for betting and suggest line ups for fantasy premier leagues. Journalists are increasingly using facts obtained from big data to corroborate their stories and often create new stories when normally none could have existed, for example stats may show a playerfis ineffective performance could be masked by the good form of the team mates around him. We will now go into a little detail of how all this is done.

2 BIG DATA IN SOCCER

2.1 Big Data in Scouting

2.1.1 Introduction. Two of the biggest commodities in soccer are the clubs and its players. As mentioned previously transfers are now some of the biggest sources of revenue for clubs. Players fetch for as high as 200 million pounds these days[13]. Also, the quest to find the next big star / the hidden gem against proven expensive players is now a mark of success. Clubs cannot freely go and sign whoever they feel are data monsters, restrictions on the number of players they can sign while at the same time the potential costs that may be involved in the transfer force clubs to make sure the investment they make are the right one. It is recommended to have a look at this article to understand what happens behind the scenes at football clubs when it comes to scouting[4].

2.1.2 Data collection. Most clubs either have acquired specific companies for scouting for example Arsenal FC acquired paid over 2 million pounds for the US company StatDNA, whose data has since been used to advise their signings. [8], and or have scouts who obtain the data themselves. As to how clubs obtain the data, most do not divulge such details to protect their strategies but consensus is that popular sites like opta which analyze matches at real time and release statistics for others to make use of[7]. Alternatively, clubs send performance analysts to feeder clubs and they track matches of prospective candidates and create data for themselves.

It is also worth noting that big data has led to only to software development but as well as hardware, for example the Adidas MiCoach a device that tracks metrics and displays it to coaches is used during training and potential scouting sessions. The article mentioned provides an example of how the device was used to realize a gem among a batch of superstars.[6]

2.1.3 Data Processing. Most articles only explain in theory how they go on about processing the data and even fewer talk about the technical aspect behind it. Corroborating from different sources [10][5][8] a general theoretical summary can be given, In the case of obtaining data from say the internet i.e., mine data from free sites like squawka, whoscored, opta. Data warehousing technologies like pig, Hadoop etc, can be used. Parsing the XML, one can store this data and applying meaningful ML algorithms with defined parameters to filter players. For example, we can mine the data for fields like chances created, distance covered etc for a league and then filter out say midfielders and chances created in order to find the next best attacking midfielder.

For clubs generating their own data, real time analysis of videos using advanced image processing technologies in tandem with their own hands on analysis they could generate data and store it again or say CSV files. These files then could be uploaded to a private databank. From these banks data warehousing can be once again performed and the previous process can be repeated.

2.2 Big Data in Training and Tactics

In todayfis world which is being driven more and more by capitalistic gains, even the worldfis most famous sport i.e., soccer cannot be spared. Sports players command huge transfer fees, MNCs are pumping billions [millions are soon becoming a thing of a past][2], and as such the even the tiniest mistake can lead to millions lost. Hence, now there is a need to augment daily operations from scouting to coaching level with technology. One of the technologies which fast invading the world of soccer is big data. One can never have enough data, data guides tactics, training session, betting, scouting and so much more. Gone are the archaic days of notes and papers and fispecialistsfi [these specialists do have a very important role to play but with the advancing times they may soon become a thing of the past]. The study looks at two crucial aspects

in soccer scouting and training, tactics. In tactics, we look at the new sensation known as fantasy leagues.

2.2.1 Introduction. Before the advent of big data, coaching was a more personalized hands-on affair, that doesnfit mean it is any less now but the amount is a lot less than before. Preparing for match involved sending scouts and making them watch the match lie and relying on their notes or analyzing videos for hours in hopes of trying to find a weak link. Coaches do spend hours in front of a TV screen but they augment it with software and now look at games from a data sided point of view, an example of this is the former coach of Everton Roberto Martinez [4]. Aside from tactics big data also is slowly invading the field of training sessions big data are being used to create customized training sessions as well as to analyze and mitigate potential injuries.

2.2.2 Data Collection. Data collection here has two aspects to its hardware and software in the previous section the software component was already discussed to a good extent. Now we will shift focus towards the hardware components and their impact/ role in data collection. Below are excerpts from the article [11] which provide excellent insight as to how data is gathered

Athletes are not only monitored by cameras in stadiums, but also by many quirky devices such as accelerometers, heart rate sensors and even local GPS-like systems. for example, the Germans in the world cup held previously in Brazil wore Adidasfi miCoach elite team system during training sessions before and during the competition.[11].The device collects and transmits information directly from the athletefis bodies, including heart rate, distance, speed, acceleration and power, and then display those metrics live on an iPad. All this information is made available live on an iPad to coaches and trainers on the sideline during training, as well as post-session for in-depth analysis. Analysis of the data can help identify the fit players from those who could use a rest.

2.2.3 Data Processing. The article [10] gives a great insight into how data obtained from devices is processed Big data is characterized using the so-called three Vfis: (1) Volume, (2) Variety and (3) Velocity. With respect to tactical analytics in soccer these concepts can be mapped in the following way:

(1) Volume refers to the size of datasets in soccer. For example, a current dataset for positional data typically encoded using Extensible Markup Language (XML) ranges between 86 and 300 megabytes (MB). Thus, storing position, event and video data from a single complete Bundesliga season results in 400 gigabytes of tracking data[10].

(2) Variety refers to different data formats and data sources. Variety can be further distinguished into (a) structured, (b) semi-structured, and (c) unstructured data. Structured data has a clearly predefined schema describing the data. In contrast, unstructured data lacks a definite schema with video data and text messages being typical examples. Semi-structured data falls in between these two extremes and consists of data which lacks a pre-defined structure but may have a variable schema[10].

(3) Velocity describes the speed with which novel data is being generated. In soccer, the velocity varies widely from real-time analysis like in the case of opta to delayed statistics released by journalists etc[10]. From this data generated Machine learning models can be applied to look for anomalies and spot out a weakness in opponents and as well as gauge areas in which the own team requires improvements.

2.3 Tools

In this section, the technological stack and possible tools for implementation are discussed.A candidate big data soccer technological stack for soccer tactics analyses should be organized along several levels.

Big data tech stack figure: 1

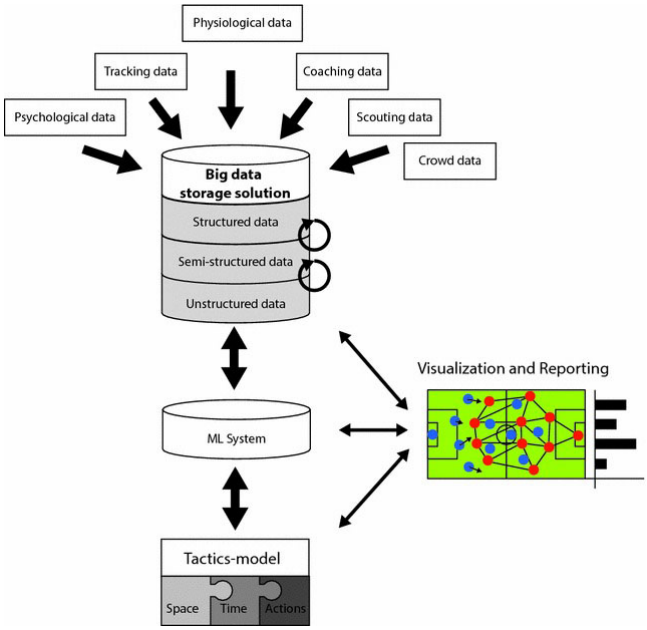

Figure 1: big data tech stack
[10]

First, the necessary infrastructure to collect the data is required. Second, a storage system is required allowing efficient data storage and access. Finally, a processing pipeline has to be established to extract relevant information from the data and to subsequently merge the information to build an explanatory and/or predictive model[10].An in-depth discussion of specific technological solutions is beyond the scope of the present study. A few useful technologies are however discussed, note we will only be discussing the software aspect since the hardware aspects have been discussed in great detail in the previous sections.

It has previously been stressed upon how difficult it is to obtain the details of technologies clubs use to run their daily operations. However, after dissecting it is not all that difficult to make an educated guess on which tools could possibly be used for the above purposes. Let's start with obtaining data in the previous sections we have already seen how opta obtains its data using live analysis[1]. Now we shall explore a new tool called Twitter Heron, which can be used to obtain information from tweets.

One of the problems with opta is that it may not cover leagues/tiers which are not profitable for it. However, football clubs generally

have very enthusiastic fan bases and with Twitter being a very convenient social media tool we can try to mine data from tweets to generate our data. Twitter heron is a real-time analytics platform developed by Twitter. It is the direct successor of Apache Storm, built to be backward compatible with Storm's topology API but with a wide array of architectural improvements. Heron supports Seamless support for different processing semantics, is efficient and scales extremely well. A good blog on why Twitter heron is ideal can be found here.[9]

Previously we touched on the subject of how scouts could use data from opta for analysis, from the internet the best way to obtain such data is to extract from XML. For this many different ways can be used. Some of the most popular manners are using Map Reduce, LogParser and even PIG.

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. The map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.[12]

Log Parser is a free command line utility for Windows that allows you to perform queries against a variety of file types including things like log files, CSV files, and XML files. This utility can even parse data sources such as the Active Directory or the Windows Event Logs. Log Parser is extremely flexible, but it is not a utility for novices. Using Log Parser requires experience with custom queries as well as with working from the command line. An example of PIG XML parsing can be found in this blog [3].We can Use Spark SQL for querying data from DBs so that it can be used to extract features and clean up data.

3 CONCLUSIONS

It can be seen the huge impact Big Data has in soccer. It has become a multi-million business. The acquisition of StatDNA by Arsenal for 2 million is proof of that. Also nowadays more and more clubs are being run entirely on big data proof of this is FC Midtjylland (Denmark) and also Brentford FC (England). Matthew Benham and Rasmus Ankersen are the pioneers in data analysis and have completely revolutionized their scouting departments. OPTA is the global leader in stats generation and is rated above 60 Million plus. Aside from just scouting potentials, it is used to shape tactics and also understand the strengths and weakness of players. While it may appear that the industry seems to have less scope of development this is only true for the top-ranked clubs. Most of the mid-table and lower league clubs still make use of traditional methods. The scope for open source software which provides a detailed scouting analysis has a huge market potential.

REFERENCES

[1] Carl Bialik. 2014. The People Tracking Every Touch, Pass And Tackle in the World Cup. *oline* 1, 1 (2014), 1. https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/

[2] Julie Cooling. na. Investing in Soccer. *na* 1, 1 (na), 1. https://www.forbes.com/sites/juliecooling/2017/03/23/investing-in-soccer/#8b404ce2ec97

[3] learnbigdataanalytics. 2000. Pig XML parsing. *online* 1, 1 (2000), 1. https://learnbigdataanalytics.wordpress.com/hadoop-eco-systems/pig/practice/xml-parsing/

[4] Tim lewis. 2012. How computer analysts took over britains top clubs. *guardian* 1, 1 (2012), 1. https://www.theguardian.com/football/2014/mar/09/premier-league-football-clubs-computer-analysts-managers-data-winning

[5] Will Luca. 20000. Data Driven Football. *online* 1, 1 (20000), 1. http://data-speaks.luca-d3.com/2017/10/data-driven-football.html

[6] Wired micoach. 2000. Big Data devices in football. *online article* 1, 1 (2000), 1. https://www.wired.com/2012/09/major-league-soccer-micoach/

[7] Optasports. 2000. generating live time data. *video* 1, 1 (2000), 1. http://www.optasports.com/about/how-we-do-it/how-we-package-the-data.aspx

[8] Outsideoftheboot. 2000. An insight into the data analysis of football. *newspaper article* 1, 1 (2000), 1. http://outsideoftheboot.com/2015/09/24/insight-into-data-analysis-in-football/

[9] Karthik Ramasamy. 2000. Why Heron? *online* 1, 1 (2000), 1. https://streaml.io/blog/why-heron

[10] Robert Rein and Daniel Memert. 2016. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *PMC* 1, 1 (2016), 1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4996805/

[11] Jure Rejec. 2000s. How Big Data is Changing the World of Football. *online* 1, 1 (2000s), 1. https://datafloq.com/read/how-big-data-is-changing-the-world-of-football/1796

[12] Tutorialspoint. 2000. map reduce definition. *online* 1, 1 (2000), 1. https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[13] Wikipedia. na. Expensive transfers. *na* 1, 1 (na), 1. https://en.wikipedia.org/wiki/List_of_most_expensive_association_football_transfers

Big Data Analytics in NCAA Football

Nsikan Udoyen
School of Informatics and Computing, Indiana University
P.O. Box 1212
Dublin, Indiana 43017-6221
nudoyen@iu.edu

ABSTRACT

This paper provides an overview of applications of big data in NCAA football by surveying current research and development work that supports the increased application of big data analytics to various aspects of NCAA football. The focus of current research is support for player performance management, injury prevention, and the use of predictive analysis to predict outcomes of games. However, the nature of interactions between players in football limit the efficacy of big data techniques in other areas such as strategy.

KEYWORDS

i523,hid342, big data, analytics, NCAA football

1 INTRODUCTION

National Collegiate Athletics Association (NCAA) football is one of the most widely watched sports in the United States. The size of the fan base and the profits that can be derived from televised games incentivize universities and other interested parties to invest in the application of big data analytics and data science methods in general to improve on-field outcomes by enabling better management of player well-being and performance. The purpose of this paper is to provide an overview of the use of data science in National Collegiate Athletics Association (NCAA) football. Recent research on the use of data science to improve various aspects of NCAA football will be surveyed, while current trends and their implications will be discussed.

2 BIG DATA ANALYTICS IN NCAA FOOTBALL

2.1 Predictive Analytics

NCAA football analysts invest a significant amount of time trying to forecast performance of various teams throughout the season. Their analysis fuels sports talk shows and other mass media programs that target dedicated fan bases, giving them a deeper understanding of the game and allowing them to learn more about their teams. Data used to support NCAA football analysts' predictions is drawn from a mix of sources such as coaches' polls, and detailed and routinely updated data on players' performance. Some of this data is combined to create composite indexes, such as ESPN's Football Power Index (FPI)[1], which are used to rank teams based on thousands of simulations of their game outcomes, and updated weekly, based on available data. Composite indexes such as the FPI support broader discussion of matchups every week, and encourage analysts to ask broader questions in previewing games, but typically are not used in any systematic way to predict outcomes.

Several researchers have applied data mining methods towards the prediction of NCAA football scores[4],[2]. Various research efforts have focused on the scope of relevant data, and how to model such data. In their paper comparing NCAA football game outcome prediction methods, Delen et al. used data on NCAA teams from 244 bowl games between 2002 and 2009 to generate and compare several predictive models[2]. They compared the performance of the models by using them to predict 2010-11 bowl game scores and found that classification-based models were better than regression-based classification methods at predicting game outcomes.

2.2 Performance Management & Player Safety

Several data mining methods have been developed to monitor athletes' performance and enable coaches to make data-driven decisions to improve results and avoid injuries. Platforms such as Microsoft's Sports Performance Platform [3] enable the collection and aggregation of biometric and other data that can be used to monitor performance. The use of wearable technology devices such as Fitbit to monitor NCAA football players has been proposed. Most efforts to apply data analytics to performance management in NCAA football focus on the evaluation and management of individual players, rather than the use of data mining to drive strategic decisions for teams during games.

In support of performance management, groups such as the NCAA Sports Science Institute gather data on injuries to college athletes and have used findings from their studies based on that data to advise the NCAA on issues such as the optimal frequency of football practices[7]. By analyzing data from the Big 12 conference, scientists at the NCAA Sports Science Institute were able to determine that the majority of injuries (and 58% of concussions) occurred during preseason practice. Their suggested guidelines, which were endorsed by 16 medical organizations, called for a reduction in the frequency of preseason practice sessions and less full-contact practice sessions.

In their paper, Ofoghi et al. describe how performance analysis requirements influence data gathering in their presentation of a general framework that applies data mining methods to sports [5]. The authors attempt to describe in their framework the most important features needed to categorize sports to enable data mining. Through their framework, Ofoghi et al. discuss the types of data that can be collected, depending on the nature of the sport being studied, and list important considerations.

Schumaker et al., list several standard data-driven metrics used to assess football teams and individual players[6]. The listed metrics include:

- *Defense-Adjusted Value Over Average (DVOA)*, which measures the success of a particular play against a defense and compares it to the average.

- *Defense-Adjusted Points Above Replacement (DPAR)*, which evaluates individual players by assessing their contribution (in points) compared to a replacement player.
- *Adjusted Line Yards (ALY)*, which assigns credit to an offensive line based on how far the ball is carried

While abundant data exists to compute the listed metrics and compare teams using them, their subjective nature makes them unreliable. DVOA, for instance, accounts for variables such as time remaining in the game, field position, and the quality of the opponent. There is no guidance on how such variables are computed or the weights assigned to each one. The ALY measures the contribution of the offensive line and the running back by rewarding the running back's individual effort for successful carries and punishing the offensive line for failed attempts. The ALY is adjusted based on league averages, which do not account for issues such as weather or bad officiating, which may have impacted a team's performance.

When used together, these metrics give a detailed view of a team's past performances. There is however, no evidence of successful use of such detailed assessments of a team's past performances to support strategic decisions during a game. The metrics are more suitable for highlighting areas of concern than predicting how well one team will fare against another before they play.

3 DISCUSSION

Research on predictive models that predict outcomes of NCAA football games illustrates the difficulty involved in capturing the nuances and complexity of the sport in a model. It also illustrates problems with the use of historical data for predictive purposes in NCAA football. For example, the data mined for the study by Delen et al., which was used to predict 2010-11 bowl games, included data points from as early as 2002, when none of the players in the 2010-11 bowl games were even eligible to play college football. It is difficult to determine how much data is sufficient to produce accurate predictions, and current data alone may not be sufficient, since some NCAA football teams may play as few as eleven games in a season.

Several features of the metrics used to describe and rate NCAA football players and teams make it difficult to use them for predictive purposes, despite the abundance of data to be collected. These include

- *The subjective nature of the metrics*
  To account for the context-specific nature of the data being gathered to describe individual and team performances, some metrics are weighted to reflect factors such as the quality of the opponent. Such subjective factors are usually not evenly considered by different evaluators, and may change as the season progresses.
- *Focus on outcome-based metrics, such as ALY*
  By relying on metrics that report only the outcomes of individual plays, data that reflects the tactics used and other technical aspects of the game are overlooked. Such metrics also ignore an opponents ability to learn and improve after a football game.
- *Inability to aggregate metrics*
  No single metric effectively describes a football team's performance well enough to enable comparison to other teams. When different metrics are combined to describe a football team's performance, the manner in which they are combined is subjective. When the metrics are combined to create a composite index used to compare teams and predict outcomes, they do not provide a complete picture of potential interactions and mismatches between teams that could influence the outcome of the game between them. A prime example of this is the Bowl College Series (BCS) formula used to select the teams that would play for the NCAA Football National Championship from 1998 to 2013.
- *Lack of context*
  When metrics are used to rate individual players, they often do not account for teammates' inputs. An example is yards-after-catch (YAC), often used by scouts to rate wide receivers. YAC reports the amount of additional yards a player gains after catching a pass from the quarterback, and should measure individual effort of the player that catches the ball. However, additional yards gained by a player after catching the ball may be due to defensive errors or assistance from teammates who block players on the opposing team. Likewise, other metrics used to rate receivers such as yard-per-catch or total yards are computed without considering the quality of the quarterback's decision-making or the defensive schemes employed by the opponent.

The use of data mining to manage player performance raises concerns over privacy and the ownership and potential misuse of the data collected[8]. The scope and amount of data collected about players has increased with the proliferation of the use of data mining methods to study player performance. In some cases, the harvesting of data collected by wearable technology devices by sportswear companies is permitted under the terms of the agreements between universities and the sportswear companies that sponsor their football teams. While companies such as Nike have stated that they have not yet begun harvesting players' biometric data, at least some of the data they could collect would not be covered by United States federal HIPA (Health Information Portability and Accountability Act) laws[9].

4 CONCLUSION

The use of data mining and analytics in NCAA football is increasing, as it has in other sports. However, due to the complexity of the game, practical uses of data analytics currently available and under exploration are in individual and team performance management and prevention of injuries. Research on data analytics, and current applications of technology to NCAA football have focused on techniques to extract meaningful information from gathered data, rather than the explanation and use of such information for predictive purposes.

The inability to account for context in data makes the use of data science to predict outcomes and influence strategy in NCAA football games difficult. The use of data primarily to compile metrics that describe past outcomes and average individual and team performance levels does not enable an understanding of their true

capabilities. There is thus a need to continue to rely on qualitative assessments by experts when making predictions or scouting individual players, and use data analytics as a supporting tool to provide relevant information to guide the discussion.

REFERENCES

[1] 2017. ESPN Football Power Index - 2017. ESPN Online. (Oct. 2017). http://www.espn.com/college-football/statistics/teamratings

[2] Dursun Delen, Douglas Cogdell, and Nihat Kasap. 2012. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting* 28 (2012), 543–552. https://doi.org/10.1016/j.ijforecast.2011.05.002

[3] Jeff Hansen. 2017. Sports Performance Platform puts data into play fi?! and action fi?! for athletes and teams. Official Microsoft Blog. (June 2017). https://blogs.microsoft.com/blog/2017/06/27/sports-performance-platform-puts-data-play-action-athletes-teams/

[4] Carson K. Leung and Kyle W. Joseph. 2014. Sports data mining: predicting results for the college football games. *Procedia Computer Science* 35, special issue of KES 2014 (2014), 710–719.

[5] Bahadorreza Ofoghi and John Zeleznikow. 2013. Data Mining in Elite Sports: A Review and a Framework. *Measurement in Physical Education and Exercise Science* (July 2013), 171–186. http://dx.doi.org/10.1080/1091367X.2013.805137

[6] Robert P. Shumaker, Osama K. Solieman, and Hsinchun Chen. 2010. *Sports Data Mining*. Springer.

[7] Jon Solomon. 2017. NCAA recommends ending two-a-day football practices and reducing tackling. CBS Sports Online. (Jan. 2017). https://www.cbssports.com/college-football/news/ncaa-recommends-ending-two-a-day-football-practices-and-reducing-tackling/

[8] Tom Taylor. 2017. Football's Next Frontier: The Battle Over Big Data. (June 2017). https://www.si.com/2017/06/27/nfl-football-next-frontier-battle-big-data-whoop-nflpa

[9] Mark Tracy. 2016. With Wearable Tech Deals, New Player Data Is Up for Grabs. The New York Times. (Sept. 2016). https://nyti.ms/2creZ4t

DevOps In Support Of Big Data Applications And Analytics

Juan Ni
Bloomington, Indiana 47401
nijuan@iu.edu

ABSTRACT

We show the relationship between devops and big data applications and analytic. The investigation will focus how devops support big data application. We Also find out the how devops on cloud computing platform impact big data application, since the carrier for most of big data application and analytic is cloud platform. At last, defining the difficulty of implement devops which is important for designing a big data application.

KEYWORDS

i523, HID 107, big data, devops, cloud platform

1 INTRODUCTION

Following with cloud computing and big data analytic prosperous, devops begain getting more and more attention. In the traditional development group, we usually have five processes which are "Analyze", "Design" , "Code", "Test", and "Maintain"[7]. According to Justin's idea, each of the traditional process is isolate with others, once one of the process group finish their part, they will simply shift the work to next group for the next level processing. I have been participated into a traditional development, once the client change their needs, traditional will take really long time to track back the work and making adjustment at traditional development team. Especially for big data application, due to the complex of big data application, the development period will be longer than usual, therefore, we need devops to help us satisfy our client needs at long period development and keep efficiency processing speed. Devops is not a set of tool, it is a methodology that include basic principle and practical. According to the concept of devops, the processes of devops are include "Code", "Build", "est", "Package", "Release", "Configure", and "Monitor" [12]. Unlike the traditional development processes, all processes of devops are connecting inside a loop, this make devops as an integration of Development and Operations.

2 THE NEED OF DEVOPS FOR BIG DATA APPLICATION

The gap of communication between development and operations on big data application is the main issue. Andrew points out that "The idea of DevOps is to tear down the silos between software developers and IT infrastructure administrators to make sure everyone is focused on a singular goal." [2] In the traditional development team, developers are not involve into the analyst activity, because the big data analyst and application developer is isolated. According to my experience, once the developer get the changing requirement from analytic team, they need to take time to understand the adjustment and reorganized the manpower inside the group. This communication delay will lower the entire processing speed, decreasing the competitiveness of big data analysis. Andre mention that "IT leaders are under increased pressure to produce results. This forces analytics scientists to revamp their algorithms. These major changes in analytic models often require drastically different infrastructure resource requirements than was originally planned for" [2] In big data application and analytic project, analyst change their algorithm ceaselessly, and the change of analytically model will make the infrastructure and resource demeaned become much different with the original one. We know that data is timeliness, big data is not a exception. If big data application processing take to much time, the outcome will be less value, so big data application need devops to prevent data losing value by fall behind.

3 THE VALUES OF DEVOPS

The main value of devops is to break down the "Wall of confusion" between developer and operator. According to Jerome's idea, devops have two main values which are fiContinuous Deliveryfi and fiBenefitsfi [6] . The "Benefits" of devops include but not limit to "Repeatability and Reliability", "Productivity", "Time to recovery", "Guarantee that infrastructure is homogeneous", "Make sure standards are respected", and "Allow developer to do lots of tasks themselves". Those benefit allow developer and operator working better as a team, and understnad each others work; according to Allerin's idea, . "Continuous Delivery" help project team decrease the application delivery period by having faster application development, high frequency update can reduce the risk and cost of changing demand during the delivery. This is extremely useful for big data application because it always requirement lot of change during delivery. The increasing of delivery frequency can let the project team more familiar with the processes of application deployment, also will getting more feedback from the user. Therefor, Haff points out the core value of devops which is "When DevOps began, so did a shorthand description for the model: It broke down the wall between dev and ops. The teams communicated better and operated with a shared set of objectives and concerns. At the extreme, there were no longer devs and ops people, but DevOps skill sets." [4].

IBM organizes the values of devops into three domains, "increase customer experience", "improve innovation ability", and "faster achieve value". [3]. According to IBM's concept, Devops is not the goal of application development, but it can let development team reach their goal. It increase customer experience via faster update, and having faster response to customer's feedback. Then we using devops to avoid rework cause by misunderstanding the demand, so the project team have time and energy to investigate new technology. Finally, once the delivery period is shorter and shorter, user can actually use the application early before the content inside the application are out date, this is important for big data application because the replacement of big data analysis algorithm is changing all the time. The above values show us that Devops is endless, it will continually develop the entire project team's technology, processes flow, teamwork, and team culture.

4 CLOUD PLATFORM AND DEVOPS

Cloud platform play a really important role in big data application designing. According to Microsoft concept, “Cloud Platform System lowers costs at all stages of the infrastructure life-cycle”[9];The expectation of cloud platform is change the capital cost to operate cost, company don’t need to figure out the cost of hardware for building a cloud server for the big data application, they can use public cloud platform really economical to prevent any waste on the computing ability. According to Allerin’s idea, “enterprises are now considering of moving their Big data and Hadoop projects to public cloud services for gaining the much-needed agility they need for their data scientists.With a scalable and flexible infrastructure platform, IT organizations and development team together can spin up virtual Hadoop or Spark clusters within minutes.” [1], so the true value of using cloud platform is decrease the barrier which slow down the development speed and developer time. Sumologic’s article describe how how automatic devops works in the some mainstream cloud platforms, according to his idea, “ The delivery pipeline collapses to a single sile where developers, testers, and operations professionals collaborate as one and as much of the deployment process as possible is automated.”[11]; most of devops work on the cloud platform can be automatic, so development team can be free from the heavy lifting of daily work such as management hardware and patch installing. Therefore, using cloud platform mean company from providing product change to providing service, which fit the big data application purpose. I think Customer won’t buy the application if the content doesn’t make sense, so let cloud platform supporting devops is good for the project team focus on the content inside the application, and be able to development more efficiently.

Mary mention that “In a true Waterfall development project, each of these represents a distinct stage of software development, and each stage generally finishes before the next one can begin.”[8],that mean user expect having high quality function and continuous update in the same time. Unlike the traditional devilment team provide “Waterfall methodology” to release event, devops provide frequent release events can satisfy customer need and maximize utilize the automation of cloud platform. Moreover, According to Nelson and Raouf’s idea,“Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications ”[10], this show the great compatibility between cloud computing platform and big data application. We know that devops are more often use for could platform. So, following the supporting by cloud platform, Devops make operate and development keep almost the same speed and flexibility which is suitable for big data business need in cloud platform.

5 THE WEAKNESS OF DEVOPS

Haff mention that“Dev teams that are making the most of this model need to focus on improved application architectures and developer workflows” [4], this is seems pretty easy but actually it has difficulty of implement, and it is the biggest weakness of devops. According to my working experience, each company have their own development tool and process, they all have their own feature, so it is really hard to have a temple that fit to every company and situation for devops. Futhermore, Jeff provided a really nice definition for the developer in devops group which is “”DevOps” is meant to denote a close collaboration and cross-pollination between what were previously purely development roles, purely operations roles, and purely QA roles. Because software needs to be released at an ever-increasing rate, the old ”waterfall” develop-test-release cycle is seen as broken. Developers must also take responsibility for the quality of the testing and release environments.” [5], this mean if a traditional developer want to fit into a devops team, that developer need to fimilar with testingfiffiapplication implementfiffiand need which almost cover every role’s job inside the team. I think a person’s energy and time is limited, if a person spend time to something, then that person spend time to other thing will be decreasing. This is same for developer, if a developer have “Many hats ”[5], then that developer probably no be able to focus on coding. And this will cause “Jack of All Trades, Master of None” [5] happen to developer, thus even the product release spedding is increasing but the quality is low. According to my experience on a big application team, for team leader, they only able to organized team and develop processes under limit source. So it is hard to having a optimal plan at most of time, for example, to let each member inside the team have same understanding and sense of duty is really hard, but the above section mention that breaking the wall between develop and operate require team member familiar with each other’s area, so the difficulty of devops is to fill the knowledge gap between each team member inside the group.

6 CONCLUSIONS

The above section show the main purpose of doing devops in big data applications and analytic is to eliminate the isolate situation between Solution Architect(big data analyst) and programmer.This can be achieve by cross education training for both architect and programmer, so they can understand the basic concept and terminology from both domain. Once they finish the training, they will suppose to have a better understanding of each others’ idea, and prevent the process off track. Furthermore, devops can let those two group testing the application environment and adjusting the foundation framework to meet the new needs, it represent faster fixing and update capacity.

7 ACKNOWLEDGEMENT

I would like to take this chance to thanks to my tutor Juliette Zerick, in process on reviewing my paper, she gave me many useful comment and advise. 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 internship supervisor who give me many idea about my topic.

REFERENCES

[1] ALLERIN . 2017. Why big data needs DevOps? (2017). https://www.allerin.com/blog/why-big-data-needs-devops [Online; accessed 19-October-2017].
[2] Andrew Froehlich . 2017. Your Big Data Strategy Needs DevOps. (2017). https://www.informationweek.com/big-data/your-big-data-strategy-needs-devops/a/d-id/1328184? [Online; accessed 19-October-2017].
[3] Gina Poole. 2017. DevOps delivers real value and successful business outcomes. (2017). https://www.ibm.com/developerworks/community/blogs/invisiblethread/entry/devops_delivers_real_value?lang=en [Online; accessed 08-October-2017].

[4] Gordon Haff . 2017. DevOps success: A new team model emerges. (2017). https://enterprisersproject.com/article/2017/6/devops-success-new-team-model-emerges [Online; accessed 20-October-2017].
[5] Jeff Knupp. 2014. How 'DevOps' is Killing the Developer. (2014). https://jeffknupp.com/blog/2014/04/15/how-devops-is-killing-the-developer/ [Online; accessed 20-October-2017].
[6] Jerome Kehrli. 2017. DevOps explained. (2017). https://www.niceideas.ch/roller2/badtrash/entry/devops-explained [Online; accessed 08-October-2017].
[7] justin . 2017. What is the Software Development Life Cycle (SDLC)? (2017). https://airbrake.io/blog/sdlc/what-is-the-software-development-life-cycle [Online; accessed 19-October-2017].
[8] Mary Lotz. 2016. Waterfall vs. Agile: Which is the Right Development Methodology for Your Project? (2016). https://www.seguetech.com/waterfall-vs-agile-methodology/ [Online; accessed 08-October-2017].
[9] Microsoft . 2017. cloud-platform-system overview. (2017). https://www.microsoft.com/en-us/cloud-platform/cloud-platform-system [Online; accessed 19-October-2017].
[10] Raouf Boutaba Nelson L. S. da Fonseca. 2015. Big Data on Clouds. *Networking, and Management* (April 2015), 2. https://doi.org/10.1002/9781119042655.ch15
[11] sumologic. 2017. DevOps as a Service. (2017). https://www.sumologic.com/devops/devops-as-a-service/ [Online; accessed 19-October-2017].
[12] Wikipedia. 2017. Devops — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/wiki/DevOps [Online; accessed 08-October-2017].

Big Data Analytics using Spark

Nisha Chandwani
Indiana University Bloomington
Bloomington, Indiana 47405
nchandwa@iu.edu

ABSTRACT

With Petabytes of data being generated every second, big data analytics has become one of the most talked about terms in the technological world. Many organizations are trying to use big data for deriving useful business insights in order to improve decision making. However, we need special tools and frameworks to analyze such large amounts of data. We discuss how big data can be efficiently analyzed using Apache Spark which is a memory based computing framework. We discuss the core components and the architecture of Spark along with its ecosystem that extends the capabilities of Hadoop MapReduce.

KEYWORDS

i523, HID203, Apache Spark, RDD, Big Data Analytics, Hadoop, MapReduce

1 INTRODUCTION

The growth of data has been following an exponential rate with huge amounts of data being generated every second. In today's world, having Terabytes or even Petabytes of data to deal with is not uncommon. The challenge lies not only in the volume of data but also in the large variance of the kind of data that has to be dealt with. This has led to the birth of one of the most talked about terms in today's technological world, i.e., Big Data. Most of the organizations today are collecting big data with the goal of extracting *value* from the exploratory analysis of this data and using this information to make business decisions. However, analyzing such enormous data is in itself a huge challenge and this is where big data analytics frameworks like Spark come to rescue. Spark is a general distributed computing framework that is optimized for in-memory processing. We show how Spark supports faster data analysis and is proving to be one of the most successful frameworks for Big Data Analytics.

2 SPARK

Spark is an open-source distributed computing framework which is based on Hadoop MapReduce algorithms [4]. However, using Hadoop MapReduce for complex tasks requires frequent disk I/O which make Hadoop less suited for low-latency tasks. To overcome this, Spark extends the capability of MapReduce by providing in-memory computing which enables it to query data much faster than disk-based engines like Hadoop [9]. Due to its memory computing capabilities, Spark is often used for iterative applications, such as Data Mining and Machine Learning [4].

Apache Spark has a well-defined architecture which is based on two main abstractions [3]:

- Resilient Distributed Datasets (RDD)
- Directed Acyclic Graph (DAG)

2.1 Resilient Distributed Datasets (RDD)

The entire framework of Spark is centered around RDD as it supports in-memory processing computation. This implies that it can store the state of memory in the form of an object across multiple jobs and the object is shared between these jobs [5]. RDD is a collection of data items that can function in parallel and is stored in memory or on disk. This parallel data computing structure is read-only and is distributed over a cluster of machines offering a restricted form of distributed shared memory. RDDs are maintained in a fault-tolerant way as the intermediate data is cached across a set of nodes. Thus, RDDs enable Spark to efficiently support iterative algorithms [7].

RDD supports two types of operations [2]:

- Transformation: Join, filter, union, map and various other operations that can be performed on existing RDDs which result in a new RDD at the end of the operation, are referred to as transformations.
- Action: Count, first, reduce and various other operations which evaluate an existing RDD and return values at the end of these operations are referred to as Actions.

2.2 Directed Acyclic Graph (DAG)

Spark consists of an advanced Directed Acyclic Graph (DAG) engine which allows programmers to develop complex, multi-step data pipeline [8]. Each Spark job creates a DAG of task stages to be executed on the cluster where each node in the DAG is an RDD partition and each edge represents a transformation to be applied on the data. This allows simple tasks to complete in a single stage whereas more complex tasks are completed in a single run of multiple stages, rather than splitting them into multiple jobs [10]. Thus, DAG abstraction eliminates the Hadoop MapReduce multi-stage execution model resulting in better performance [3].

3 SPARK ARCHITECTURE AND HARDWARE INTRODUCTION

Spark is built in programming language Scala and is run on Java Virtual Machine (JVM). In addition to Scala, it provides API for Java and Python as well. For running an application, Spark provides the following two options [10]:

- Interpreter in the Scala language distribution allows users to execute their queries on large data sets through Spark engine.
- Users can write their applications as Scala programs called driver programs. These driver programs can be then compiled and submitted to the cluster's master node.

Apache Spark uses a master/worker architecture as shown in Figure 1. It mainly consists of a driver program (SparkContext),

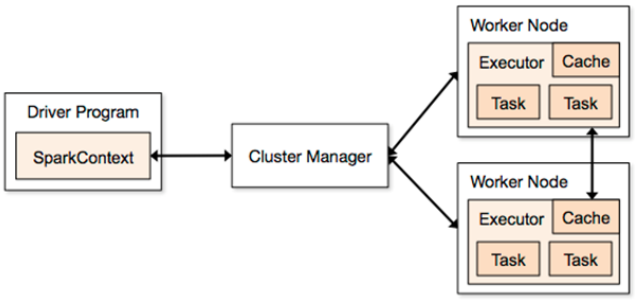

Figure 1: Spark Architecture [1]

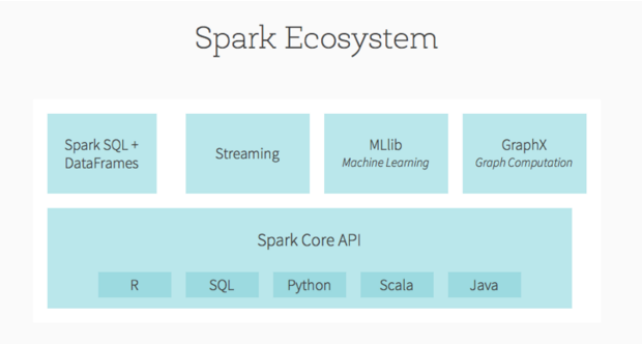

Figure 2: Spark Ecosystem [6]

workers (executors) and a cluster manager which are described below [3]:

- Driver Program: This program executes the main function of the Spark application. It is also responsible for the creation of the SparkContext object which basically coordinates the independent sets of processes running for an application on the cluster. The main components of the driver program are - DAGScheduler, TaskScheduler, BackendScheduler and BlockManager that translate the user code into Spark jobs that are executed on the cluster.
- Executor: These are the worker processes that are responsible for the execution of tasks sent by the SparkContext object. Some of these tasks include processing the data, reading from and writing data to external sources, performing computations and storing the results in in-memory cache or on hard disk drives.
- Cluster Manager: This is an external service that is responsible for obtaining resources on the Spark cluster and distributing them to the Spark jobs.

Being a memory-based computing platform, one of the most important factors of the Spark cluster is the memory. All the nodes, i.e., the driver and the executor nodes, should be equipped with at least 8 GB of memory for Spark to run well. For the cluster manager, Spark currently supports the below three deployments [4]:

- Standalone: It is a simple cluster manager included with Spark. Since Spark Standalone is available in the default configuration, it is the easiest way to set up a cluster and run applications on Spark.
- Apache Mesos: It is a general cluster manager that provides API for resource management and task scheduling across multiple nodes
- Hadoop YARN: It is the resource manager in Hadoop 2 which was added to Spark in version 0.6.

4 SPARK FOR BIG DATA ANALYTICS

With a large number of companies now looking to expand their advanced analytics capabilities, the ecosystem of Spark is right out of the box, making advanced analytics a reality. This ecosystem, as shown in Figure 2, provides an impressive set of high-level tools which include - Spark SQL for SQL, MLlib for machine learning, GraphX for graph processing and Spark Streaming [4]. Each of these components is-

- Spark Core API is the foundation of the overall ecosystem that provides task scheduling, dispatching and basic I/O functionalities. It is available through API in languages like Java, Python, Scala and R.
- Spark SQL is Apache's Spark module for supporting SQL implementation. It provides seamless integration of SQL queries with Spark programs. It provides a common way to connect to a variety of data sources such as Hive, JSON, JDBC, etc.
- Spark Streaming is an extension to the core Spark API which provides the capability to process streaming jobs along with batch jobs. The languages supported are Java, Python and Scala.
- MLlib is Spark's scalable Machine Learning library which is usable in Java, Python, Scala and R. MLlib supports high-quality algorithms such as classification, regression, clustering, recommendation, dimensionality reduction, etc. MLlib leverages Spark's excellence in iterative computing enabling it to run faster than MapReduce on huge datasets.
- GraphX is Spark's parallel computation API that is used for charts and graphs processing [4]. GraphX extends the capabilities of Spark RDD by introducing RDD graph which is a directed multi-graph with properties connected to each node and the edge [8].

One of the challenges of analyzing big data is that it can come in any shape or size. Thus, whether big data is to be processed offline (Spark Core) or on the fly (Spark Streaming), whether it is structured (Spark SQL) or connected in nature (GraphX), Spark ecosystem is a framework that can be used extensively in big data analytics.

5 SPARK VERSUS HADOOP MAPREDUCE

Both Spark and Hadoop MapReduce are extensively used in big data analytics, however, Spark has some major use cases over Hadoop [10]:

- Unlike Hadoop, Spark supports interactive data mining and data processing
- Spark outperforms Hadoop when it comes to iterative algorithms in machine learning as it keeps working sets in memory for efficient reuse

- Spark supports efficient stream processing which is one of the major advantages over Hadoop
- Spark is faster than MapReduce in execution

Though Spark has many advantages, Hadoop MapReduce can prove to be more efficient when it comes to batch processing for data with size greater than the available memory.

6 CONCLUSION

With the increasing volume of data, big data analytics is only going to become more critical for businesses decisions. Analyzing data at a huge scale presents many challenges and as we showed, Apache Spark can be very useful in overcoming these challenges. Over past few years, though Hadoop MapReduce has been one of the prime big data analytics framework, we showed how Spark has some major use cases over Hadoop. Though Apache Spark is a relatively young data project, it has already been adopted by a wide range of industries for big data analytics. We provided an introduction to Spark and discussed its architecture and the core components. As future work, we can discuss a case study and show how Spark processes big data in a more efficient manner than Hadoop MapReduce.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Apache Spark. 2017. Cluster Mode Overview. (2017). http://spark.apache.org/docs/1.3.0/cluster-overview.html

[2] DeZyre. 2016. Apache Spark Ecosystem and Spark Components. (02 2016). https://www.dezyre.com/article/apache-spark-ecosystem-and-spark-components/219

[3] DeZyre. 2017. Apache Spark Architecture Explained in Detail. (03 2017). https://www.dezyre.com/article/apache-spark-architecture-explained-in-detail/338

[4] Jian Fu, Junwei Sun, and Kaiyuan Wang. 2016. SPARK–A Big Data Processing Platform for Machine Learning. In *Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), 2016 International Conference on*. IEEE, IEEE, Wuhan, China, 48–51.

[5] V Srinivas Jonnalagadda, P Srikanth, Krishnamachari Thumati, and Sri Hari Nallamala. 2016. A Review Study of Apache Spark in Big Data Processing. *International Journal of Computer Science Trends and Technology* 4, 3 (2016).

[6] KDnuggets. 2016. Top Spark Ecosystem Projects. (03 2016). http://www.kdnuggets.com/2016/03/top-spark-ecosystem-projects.html

[7] Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu, and María S Pérez-Hernández. 2016. Spark versus flink: Understanding performance in big data analytics frameworks. In *Cluster Computing (CLUSTER), 2016 IEEE International Conference on*. IEEE, IEEE, Taipei, Taiwan, 433–442.

[8] Srini Penchikala. 2015. Big Data Processing with Apache Spark - Part 1. (01 2015). https://www.infoq.com/articles/apache-spark-introduction

[9] Abdul Ghaffar Shoro and Tariq Rahim Soomro. 2015. Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology* 15, 1 (2015).

[10] Ankush Verma, Ashik Hussain Mansuri, and Neelesh Jain. 2016. Big data management processing with Hadoop MapReduce and spark technology: A comparison. In *Colossal Data Analysis and Networking (CDAN), Symposium on*. IEEE, IEEE, Indore, India, 1–4.

Big Data Analytics and High Performance Computing

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

ABSTRACT

This paper provides an introduction to Big Data and High Performance Computing and tries to find how they are related to each other. We describe what exactly is Big Data and High Performance Computing. We then describe what technologies are in use in these respective fields and technology that can be used to combine them.

KEYWORDS

i523, hid204, Big data, High Performance Computing, SPIDAL

1 INTRODUCTION

Data is growing faster than ever, and at the same time, it is becoming obsolete faster than ever. The challenge is to how quickly and effectively one can analyze the data and gain insights that can be useful to solve problems. High Performance Computing plays an important role in running predictive analytics, especially when time is of crucial importance. In this paper, we analyze the ecosystem of the two data-intensive applications. We discuss the important features of the two fields, and then compare the functionality of the two paradigms.

2 BIG DATA

The quantity of computer data generated is growing exponentially in this world for many reasons. Retailers are building vast databases of recorded customer activities. Organizations working in logistics, financial services and health-care are also capturing more data. Social media is creating vast quantities of digital material. Big data is a term used for a combination of structured and unstructured data which has a potential to be mined for information [8]. It is often characterized by 3Vs : the enormous **Volume** of data, the **Variety** of data and the **Velocity** at which data is processed.

Here, Volume poses both the greatest challenge and the greatest opportunity as big data could help many organizations to understand people better and allocate resources more effectively. Big Data velocity also raises a number of issues as the rate at which data is flowing into many organizations is exceeding the capacity of their IT systems. In addition, user increasingly demand data to be streamed to them in real-time and delivering this can prove quite a challenge. Finally, the variety of data-types to be processed are becoming increasingly diverse. Today not only text documents, but audio, video , photographs are all equally important source of data [8].

Recently Big data has been connected with terms such as data analytics, predictive analytics or any other kind of analytics which helps an organization to predict the user behavior so that they can improve their business. Data sets have been growing so rapidly mainly due to increasing number of ways data can be collected such as smartphones, your internet history or even your search history on a website.

3 HIGH PERFORMANCE COMPUTING

High-Performance Computing (HPC) is the use of parallel processing for running applications efficiently and quickly [6]. This term is especially used for computing architecture having capacity of more than a teraflop operations per second. It involves a lot of distinct computer processors working together on a complex problem. The complex problem is divided into smaller parts and distributed among the processors which are inter-connected using an architecture which is either massive centralized parallelism, massive distributed parallelism or something else entirely.

Massive Centralized Parallel computing refers to a computer architecture in which several high processing nodes are connected via a fast local area network. All these pseudo independent nodes are coordinated by a central scheduler. All the processors are connected to a single piece of memory. It is essentially a bigger version of a multi-core processor. It used to be the most common type of HPC architecture 15 years ago, but we do not see much of them anymore. This type of architecture is quite expensive and does not really scale [3].

Massive Distributed Parallel computing refers to a computer architecture in which several high processing nodes are interconnected but with a more diverse administrative domain. It is a more opportunity based architecture in which the resources are allocated on the basis of their availability instead of having a centralized scheduler. The way these different nodes communicate with each other is standardized through a library called Message Passing Interface(MPI) [3].

Almost every Super Computer these days is a hybrid of Distributed and Shared memory in some way. Each node will be a shared-memory system. The network connecting these nodes will be some sort of topology. Along with the architecture, the way code is written needs to get optimized as well. Parallel computing is the key to increase the performance of Super Computing. Ideally, if you have T processors, you would like your program to be T times faster. But that is not the case. This is because not all parts of a program can be successfully split into T parts which can be processed in parallel. Splitting up the program might even cause additional overheads such as communication.

HPC is typically used for scientific research or simulation and analysis of an environment through computer modelling. HPC brings together several computer technologies such as Computer Architecture, algorithms together to solve these high process demanding problems.

4 BIG DATA AND HIGH PERFORMANCE COMPUTING SOLUTIONS

4.1 Amazon Web Services

Amazon Web Services(AWS) provides a variety of tools which are not only capable of handling huge amount of data but also provides technology and techniques for working productively with data at any scale. Another advantage of using AWS for big data analytics is the low cost at which amazon provides these tools. There is no capital investment required, no subscription requirements. Along with this, the ease with which you can configure these services is incredible. Anyone with a basic knowledge of command line can configure these tools with ease. Some of the major analytics tools provided by AWS are Amazon S3, Amazon Kinesis, Amazon DynamoDB, Amazon RedShift and Amazon Elastic MapReduce.
Amazon S3 is an object storage built to store and retrieve any amount of data from anywhere such as web sites and mobile apps, corporate applications etc. It is the only cloud storage solution with query-in-place functionality, allowing you to run powerful analytics directly on your data at rest in S3.[1] Amazon Kinesis is real-time streaming and processing for BigData. It is a highly-durable buffer that can handle all that work-load on the front-end as well as on the back-end with the help of series of EMR nodes which can give you an almost realtime analytics [2].

Amazon DynamoDB is a NoSQL Database with high throughput and low latency for both read and write operations. It is a fully managed cloud database and supports both document and key-value store models. Amazon RedShift is a petabyte scale data warehouse which is massively parallel with over 1000 nodes running at a time.

4.2 Apache Hadoop Framework

The most widely known technology that helps to handle large-data would be a distribution data process framework is Apache Hadoop. It is an open-source framework used for processing huge datasets using a Map-Reduce model. It is based on a master-slave architecture where low-end commodity hardware is interconnected using ethernet. The framework broadly consists of 2 components, the storage part known as Hadoop Distributed File System(HDFS), and the processing part known as Map-Reduce [9]. The Master node split large files into smaller parts and distributes them across the slave nodes. After this, it sends the same code to every node which is used to process the data.
In the Map step, the slave nodes applies the map function to the data and stores the output temporarily. In the Shuffle step, slave nodes reshuffle data between them on the basis of key-values pairs such that data belonging to particular key is located on the same node. After this, slave nodes work process the respective keys in parallel. This results in increased efficiency as all the nodes are working in parallel independently. In the end, the MapReduce system collects the Reduce output from each node and combines it to produce final result.
MapReduce is useful in a wide range of applications, including distributed pattern-based searching, distributed sorting, web link-graph reversal, Singular Value Decomposition(SVD) and other Machine Learning algorithms [9].

4.3 Hybrid of Hadoop and HPC

There has been convergence at many levels between HPC and Hadoop even though they were originally created to fulfill completely different purpose. HPC was designed for high-end, parallel computing jobs whereas hadoop was designed for cheap data storage and computing jobs.
There has been research going on offering a scale of comparison for different data-intensive computing fields, including blending the “best of both” computing paradigms using a hybrid of MPI and Hadoop. “The goal is to successfully bring the two data-intensive computing paradigms together to share the developments versus “reinvent the wheel” on either side” [7]. Machine Learning is another area which will have a lot to gain by this hybrid of HPC and Big Data as most of the ML algorithms are based on Linear Algebra which is a common HPC problem. if we run K-Means on MPI and Hadoop, MPI gave out better results than Hadoop. But the second generation Hadoop frameworks such as Spark gave out significantly better performance as they are adopting techniques such as effective collective operations which were previously only found in HPC architecture [7].
Another approach that has been proposed to converge these 2 systems is running Hadoop on top of HPC. However, a lot positives of Hadoop such as higher cluster Utilization are lost in this approach. Furthermore, Hadoop2(YARN) is capable of implementing both HPC applications and data- intensive applications but it still needs work [5].

4.4 Scalable Parallel Interoperable Data-Analytics Library (SPIDAL)

Many of the currently available commercial environments are more shifted towards the data-intensive paradigm. To make these environments work with HPC, there is need for HPC to look towards JAVA to run its codes as most of these commercial environments use JAVA whereas HPC has traditionally preferred C,C++. In the last few years, development has been done in this domain and SPIDAL JAVA has demonstrated significant performance gains when running on clusters upto 3072 cores [4].The developer friendly Java interface in SPIDAL Java will help to integrate it with other big data platforms such as Apache Hadoop, Spark, and Storm in future [4].

5 CONCLUSIONS

Big Data Analytics and High Performance Computing are quite similar paradigms even though they were built for completely different purpose. In the next few years, it is not unrealistic to believe that hadoop jobs to be processed on high end super computers instead of low end commodity infrastructure it presently runs on. This will not only help the Big Data industry but also other fields such as Machine Learning which certainly requires high end computing architecture.

ACKNOWLEDGMENTS

The author 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 this paper.

REFERENCES

[1] Amazon. 2017. Big Data Analytics and High Performance Computing. Web Page. (Oct. 2017). https://aws.amazon.com/S3/ HID: 204.

[2] Amazon. 2017. Big Data Analytics and High Performance Computing. Web Page. (Oct. 2017). https://aws.amazon.com/kinesis/ HID: 204.

[3] Ira Brodsky. 2017. Scale grid computing down to size. Web Page. (Oct. 2017). https://www.networkworld.com/article/2339444/software/scale-grid-computing-down-to-size.html HID: 204.

[4] Saliya Ekanayake, Supun Kamburugamuve, and Geoffrey Fox. 2016. SPIDAL Java: high performance data analytics with Java and MPI on large multicore HPC clusters. (04 2016), 3 pages.

[5] Shantenu Jha, Judy Qiu, André Luckow, Pradeep Kumar Mantha, and Geoffrey Charles Fox. 2014. A Tale of Two Data-Intensive Paradigms: Applications, Abstractions, and Architectures. *CoRR* abs/1403.1528 (2014). http://arxiv.org/abs/1403.1528

[6] J.A.N. Lee and J. Impagliazzo. 2004. *History of Computing in Education: IFIP 18th World Computer Congress, TC3 / TC9 1st Conference on the History of Computing in Education 22–27 August 2004 Toulouse, France*. Springer US. https://books.google.com/books?id=J46GinHakmkC

[7] André Luckow, Mark Santcroos, Ole Weidner, Ashley Zebrowski, and Shantenu Jha. 2013. Pilot-Data: An Abstraction for Distributed Data. *CoRR* abs/1301.6228 (2013). http://arxiv.org/abs/1301.6228

[8] Wikipedia. 2016. Big Data Analytics and High Performance Computing. Web Page. (June 2016). https://en.wikipedia.org/wiki/Big_data HID: 204.

[9] Wikipedia. 2016. Big Data Analytics and High Performance Computing. Web Page. (June 2016). https://en.wikipedia.org/wiki/Apache_Hadoop HID: 204.

The Internet of Things and Big Data Analytics

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

ABSTRACT

The Internet of Things, or IoT, is all about data from connected devices. Millions of consumer and industrial devices drive IoT growth and challenge with data volume and variety. Big Data Analytics helps combing through these high volumes of complex IoT data into meaningful business insights.

KEYWORDS

i523, hid306, Internet of Things, IoT, Smart Devices, Sensors, Big Data Analytics

1 INTRODUCTION

The Internet of Things (*IoT*) is the collection of devices having sensors, actuators, and Internet connectivity to gather, send and receive data. Devices of all types: cars, thermostats, implants for radio-frequency identification (RFID), pacemakers and more - have become smarter, opening up the need for their connectivity with the Internet. Most of the IoT activity is centered in transportation, smart environments, manufacturing, and consumer applications like wearable gadgets, but within three to five years all industries will roll-out some kinds of IoT initiatives. *Gartner, Inc.* predicts that 8.4 billion IoT devices will be in use by the end of 2017 and will reach 20.4 billion by 2020 [1].

2 IOT EXAMPLES

The rise of IoT changes everything by enabling *smart* things. Products and environments are becoming smarter. Broadly speaking, two kinds of IoT are emerging: *Consumer IoT* and *Industrial IoT* [9]. Products such as Apple Watch, Fitbit, Smart TV, etc. are considered Consumer IoT. Examples of Industrial IoT are: manufacturing equipment and medical devices. A few more examples of IoT include:

2.1 Smartphones

Modern Smartphone is a good example of IoT device as it has sensors (GPS, compass, proximity, accelerometer, etc) and can connect to Intenet using Wi-Fi or Cell data. It can monitor location, workouts and movements throughout the day.

2.2 Smart Homes

Connected devices like security system, garage opener, thermostat and refrigerator need to communicate with each other to make the home as Smart Home. As an example, once the user reaches home, his car can communicate with the garage opener to open the garage and the main door can unlock automatically in response to the app installed on the Smartphone. The thermostat can adjust the temperature due to sensing his proximity. Refrigerator can reorder groceries using built-in scanning, sensors and actuators.

2.3 Smart Cities

Automated transportation with smart traffic lights, use of road sensors and smart parking, smarter energy consumption using smart grids and smart meters, environmental monitoring with usage of Wireless Sensor Network (WSN) are all examples of IoT applications for smart cities.

2.4 Smart Medical Alerts

Proteus invented the *smart pill* that has a sensor for tracking cardiometabolic conditions of the patient. Once the patient takes this medication, the sensor starts sending signals to a patch worn on the skin, which logs patient diagnostic information and other metrics like activity patterns to the app of patience's Smartphone. Supervised doctor can also be notified about the patient with the details [7].

2.5 Smart Aircrafts

Rolls-Royce is building *Smart Aircrafts* to track engine performance, fuel usage, air traffic, routing details and weather conditions using sensors. The data from these sensors can be analyzed for improvements of the design of aircraft engines [5].

3 NEED OF BIG DATA

The true value of IoT is not in just the Internet-connected devices; the value lies in making context-aware relevant data and converting the result to enterprise-grade, tangible and *actionable* business insights. The IoT and Big Data are highly interrelated: millions of Internet-connected devices will generate high volumes of data. As devices (*things*) turn more digital, IoT will analyze complex data-structures, and respond intelligently in real time.

Big Data is defined by *four Vs*: volume, variety, velocity and veracity [4]. (a) Volume: Companies collect large amounts of data including transactions, sensor data and social media, and store them for later processing. (b) Variety: Data comes in various formats: structured, semi-structured and unstructured. Structured data usually come from RDBMS systems. Audio, video, binary and text documents are examples of semi-structured and unstructured data. Traditional relational databases (RDBMSs) will not be suitable for scale out distributed processing to handle such volume and variety. Alternatives like *Hadoop ecosystem*, with Distributed File System, Map, Reduce, etc. aspects, allows complex data processing. (c) Velocity: Data can come in batches, near-real time and real-time. Sensor data from medical devices might need immediate processing. (d) Veracity: Big Data Veracity refers to the noise and outlier data.

Data mining will address these concerns using *data cleansing* and *normalization* techniques.

4 IOT BUILDING BLOCKS

To scale the needs of IoT, the strategy should include infrastructure and applications that process and leverage machine and sensor data accordingly. At the moment, IoT platforms are often custom-built functional architecture. Enterprises that take the first step into this new market should look for interoperability between existing systems and a new IoT operating environment. The building blocks of an ideal IoT platform include:

4.1 Sensors and Actuators

A major part of the IoT is not so much about smart things (devices), but about sensors and actuators. Smartphone would not have been smarter if it does not have an array of sensors embedded in it. A typical smartphone is equipped with five to nine sensors, depending on the model. Both *Sensors* and *actuators* are types of transducers which convert energy from one form to another, whereas sensors, are mainly applicable at the input and actuators take part in the output of a smart device [3]. Sensors detect, quantify and convert recognized signal such as variations in pressure, heat or brightness into an analog or a digital electrical output that can easily be read and process. Thermometer senses and quantifies temperature into digital readable format, hence it is a sensor. Actuators are mechanical devices, such as switches, which produce signals by mechanical means. There are many types of sensors with endless capabilities to handle various use cases of IoT ranging from simple consumer to advanced industrial scenarios.

4.2 Network Connectivity

Wi-Fi and Cellular: 3G/LTE/4G are the most common network connectivity options for smart devices. Bluetooth and Zigbee are popular for short-range network communication. Thread technology is aimed at home automation applications and TV White Space, unused TV buffer channels, are providing broadband Internet access for wider area IoT-based use cases. Factors such as range, power usage, security and life of the battery will dictate the choice of which networking technologies to use. In March 2015, the Internet Architecture Board that oversees the technical evolution of the Internet released a guide to IoT networking. This outlined four standard communication models used by IoT smart devices: Device-to-Device, Device-to-Cloud, Device-to-Gateway, and Back-End Data-Sharing [6].

4.3 Collaboration and Security

Human and organizational behavior is critical in realizing the value of IoT approaches, and it is particularly important in shifting an organization to demonstrate clearly what will change, how it affects people, and what they stand to gain from IoT applications. Tons of collected IoT data could easily contain sensitive information about people and operations, and can even lose control of critical systems. Beyond protecting personal privacy and business secrets, as more systems become automated, the risk of attacks becomes both more likely and more impactful.

Devices themselves should be secured, as should operating systems, networks and every other exposed piece of technology along the way. The roles of users, administrators and managers should be individually defined with appropriate access and strong authentication embedded in the design. A multi-layered approach to security is essential, and it should have checks and balances to reinforce protection and, if necessary, diagnose any breaches. For the IoT to work effectively, all the challenges around regulatory, legal, privacy and cybersecurity must be addressed; there needs to be a framework to exchange data securely over wired or wireless networks across devices. To address these challenges and for better IoT interoperability, one key player, *OneM2M* published Release 1, with 10 specifications covering requirements, architecture, security aspects, Application Programming Interface (API) specifications and mapping them to industry scenarios [2].

4.4 Cloud Computing

The cloud computing brings needed agility, scalability, storage, processing, global reach and reliability to an IoT platform. Flexible scalability can be achieved by using (a) Cloud Centric IoT: Good choice for low-cost things where data can easily be moved, with few ramifications (b) Edge Analytics: Ideal for things producing large volumes of data that are difficult, costly or sensitive to move, and (c) Distributed Mesh Computing: *Future-ready* multi-party devices automatically collaborate with privacy intact [8].

4.5 Big Data Analytics

Big Data Analytics, in the context of IoT, mainly refers to diagnostics, predictive maintenance, anomaly detection and reliability analysis using statistical tools and techniques with business acumen to explore hidden information from the sensor data. It applies data mining and machine learning algorithms to volume of data coming from multiple sources with various types of data formats. Typical data analytics workflow include: gathering structured and unstructured data, cleaning the data before modeling, evaluating and visualizing to make them usable for business decisions. Data modeling is, the heart of analytics, to better understand, quantify using statistical algorithms and then visualize the model to comply to the business context. Exploratory data analysis and predictive analytics are two major groups of tasks in the data modeling. Exploratory data analysis uses various techniques to provide useful textual and visual summaries of the characteristics of the data. Predictive analytics focuses on classification and numerical regression tasks.

5 CONCLUSION

The Connected Devices, also knows as the Internet of Things, are influencing people and corporations for ultimate functionality and superior usage of the Internet. Almost all the devices are or will be getting connected to the Internet. The goal of a connected IoT ecosystem is to get the most out of the Internet of your things in your context. Industrial IoT side, it is becoming disruptive yet inevitable for companies to welcome it. Creating a connected IoT ecosystem that maximizes business value, collaboration is needed with technologies, data, process, insight and people. However, security and privacy will continue to be the key concerns to IoT

growth. Innovative organizations are starting to address these concerns and pushing IoT devices to use today.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Garner. 2017. Garner Press Release. (Feb. 2017). http://www.gartner.com/newsroom/id/3598917

[2] IoT Interoperability. 2015. IoT Interoperability. (Jan. 2015). http://www.onem2m.org/images/files/oneM2M-whitepaper-January-2015.pdf

[3] Adrian McEwen and Hakim Cassimally. 2014. *Designing the Internet of Things*. Wiley, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom.

[4] Charles McLellan. 2015. The internet of things and big data: Unlocking the power. (March 2015). http://www.zdnet.com/article/the-internet-of-things-and-big-data-unlocking-the-power/

[5] Rene Millman. 2016. Rolls Royce, Microsoft team up to put IoT in the air. (April 2016). https://internetofbusiness.com/rolls-royce-microsoft-team-put-iot-air/

[6] Karen Rose, Scott Eldridge, and Lyman Chapin. 2015. *The Internet of Things: An Overview*. Technical Report. Internet Society. https://www.internetsociety.org/resources/doc/2015/iot-overview

[7] Adam Rubenfire. 2016. *Smart-pill startup Proteus eyes cardiometabolic drugs*. Technical Report. Proteus. http://www.modernhealthcare.com/article/20160404/NEWS/160409967

[8] Natalia Vassilieva. 2016. *Sense Making in an IOT World: Sensor Data Analysis with Deep Learning*. Technical Report. HP Labs. http://on-demand.gputechconf.com/gtc/2016/presentation/s6773-natalia-vassilieva-sensor-data-analysis.pdf

[9] Guest Writer. 2017. Consumer IoT vs. Industrial IoT - What are the differences? (July 2017). https://www.iotforall.com/consumer-iot-vs-industrial-iot/

BigData Analytics using Apache Spark in Social Media

Lokesh Dubey
Indiana University
3209 E 10th St
Bloomington, Indiana 47408
ldubey@indiana.edu

ABSTRACT

Social Media, as organic and diverse it is, is also a vital source of very large amount of data. And it increased even more with the introduction of Smart Phones. As it has been established very well in recent years that Social Media and the data derived from it not only helps with decision making for substantial businesses but also helps considerably for marketing and increasing business revenues. We explore various benefits and techniques of using Big Data technology Apache Spark in unison with enormous Social Media data and how it can overcome the shortcomings of Traditional Analytics Technologies. We illustrate application of Spark with Social Media data with a few Social Media use cases pertaining to product enhancement and marketing.

KEYWORDS

i523, hid309, Apache Spark, Hadoop, Social Media Analytics, Marketing, Cloud Computing

1 DATA

1.1 Traditional Data

Criticality of data has been an accepted fact right from the beginning of computing world. In fact, when the first computer was invented the first few operations and features provided by first generation computers where simple file creations, saving the data and performing calculations on them. Since then types of data and size of data has come a long way along with the advancement in the technology. However, before the introduction of Social Media to the computing world, traditional data typically remained highly structured, static and rigid [11]. A substantial part of traditional data was generated and handled in Banking, Health and Insurance domains. But most of this data stayed extremely monotonous, relational in other words structured and rigid. These data types were always constant, brittle and it was very easy to assess their growth if any in future. Because of which it was easy to forecast what kind of infrastructure and technology needed to be procured.

1.2 Social Media Data

In recent years social media has proliferated at such an exponential rate that the sheer amount of data that is being generated is becoming a challenge for traditional technologies to handle. Initially social media was, as the name suggests, medium for socializing. And the primary focus of social media was upon social interactions of humans on a digital platform which helped with fast progressing life style where the frequency of physical social interactions was reducing day by day. Social media sites, like Facebook, Twitter etc., became extremely popular at least within young generation. It started to become an extremely simple way to socialize, catch up with friends, and sharing life events with others merely by login on those sites on internet with the luxury of not moving physically anywhere and save resources and time. And of course internet's vast reach and speed made it a very likable and a viable solution. This, in effect became a huge source of data generation. Every social media user, logging on to a social media site, sharing his own information in form of photos, videos and text, and not just that, a user liking, viewing others photos, videos, status shares, became a huge source of data generation [6]. Computing world was vary of this vital change and looked at this immense amount data as a viable source for gathering different statistics of different demographics [3].

However, with the introduction of Smart Phones the whole paradigm of social media changed [1]. Now, rather than waiting for getting an internet access on a desk to visit social sites, a user had access to all these social sites on his hands. Which essentially provided a way to socialize, share and grasp, all the information from friends and other public information through out the day [14]. This major paradigm shift in social media not only increased the amount of data that was being generated but also provided various other perspectives on how better this data can be used. The data that is being generated by Social Media is used in multitude of domains with a variety of motives [15]. Data sources can be Streaming APIs, where data is being provided almost in real time, simple REST APIs to retrieve data and possibly files archived on file servers to be consumed. Data formats can be comma or any delimiter separated files, JSON[1] files, html etc.

In addition to the wide variety of data sources and formats what can be mined from this data is also very diverse [15] [12]. Commercially, this data can be used to improve on the products by mining for constructive feedback for the productions and the same data can be used in marketing for increasing sales and driving the decision making process. But there are endless possibilities of using this immense amount of data for other analysis. For example, early detection and tracking of diseases and epidemics [19].

2 BIG DATA PROCESSING

2.1 Traditional Analytics Methodologies, Challenges

Data and specifically Big Data has been around for some time. However, the data has almost always have been structured. There have been a lot of work done in the field of data warehousing and there are some other traditional appliance based warehousing

[1] In computing, JavaScript Object Notation or JSON, is an open-standard file format that uses human-readable text to transmit data objects

systems like Netezza[2], Teradata[3] which are also used for a lot of analytics. These systems however have their own limitations and if not all, they do not perform well on the contemporary Social Media data [16] when the objective is to handle complete data in real time. There are some explicit and implicit problems using these traditional technologies and methodologies with Social Media data. Explicit problems are the type of data. There are multiple data sources, formats and types in social media which are difficult to be incorporated in these traditional systems. For example, the data sources could be a stream of twitter, unstructured live chat data from a chat server, various formats of data like JSON, comma separated. These data types can very well be integrated within these systems as well but there's a huge cost to massage and transform the data to be made usable by these traditional systems. Other than the explicit challenges there are some implicit challenges which are faced when trying to ingest and processing data for which the size and its frequency is not fixed. In traditional technologies like Netezza, Teradata we have to understand our data first not only on the structure but also on the size of the data before hand so that appropriate capacity on the appliance can be procured. But with Social Media data, which can be of any type, format, size, its difficult to scale the traditional systems this quickly [11]. Because of these challenges the traditional analytics systems are not completely obsolete as there are still a lot of other data sources other than social media but for Social Media specifically when our concern mostly tackling this immense amount of data its better to move towards a technology which can handle any sources of data, formats and types of data which can be achieved very easily with a technology based off Cloud Computing. With these challenges, the traditional data methodologies face some limitations, which are summarized by Krishnan with the following sentence 'Lack of scalability due to processing complexities coupled with inherent data issues and limitations of the underlying hardware, application software, and other infrastructure' [13].

2.2 Cloud Computing

As explained in traditional analytics methodologies and traditional data before one of the major challenges in handling the ever growing and dynamic data was being able to foresee the amount of data that needs to be processed and to be able to estimate the amount of hardware/infrastructure to be procured. Both of these problems couldn't be solved by traditional warehousing and on premise or even off premise labs with high performance infrastructure. Because these systems are not scalable to the needs of big data. As far as the infrastructure for Big Data is concerned introduction of Cloud Computing was a ground breaking advancement which opened up the doors for numerous possibilities [13]. With on demand computing and on domain scale up, scale down features which were provided by a 3rd party Infrastructure as a Service (IaaS) service providers it was extremely easy to manage the dynamic data. With Virtualization, in Cloud Computing, big Infrastructure providers take care of all of the infrastructure needs and provide on demand service to provide high configuration and high performance virtual machines on demand, which can also be backed up passively in form of snapshots and can be recovered back to avoid any kind of data or infrastructure loss. These features are seldom available in traditional on premise infrastructure and if it is then it comes with a very high cost. From cost point of view as well these machines can be purchased on hourly billing rates and the user only pays for the time the machine was used. Other than compute (Memory and CPU) advancements were made on making storage highly scalable, fast and manageable like the vms in form of SAN[4] Storage with very high IOPS and Object Storage[5] for providing highly reliable and easy and remotely accessible data storage for huge data archival or even for using the same storage for Big data I/O even over network [10]. In last decade, Cloud computing grew much more than just being IaaS providers and various other providers used IaaS underneath and started providing Platform As As Service and eventually Software As A Service. It is explained later in more detail but Cloud Computing has progressed enough to even provide MapReduce and Hadoop platforms as a service.

2.3 Hadoop

Biggest breakthrough in the field of Big data were the two research paper released by Google Inc 'The Google File System' [8] and 'MapReduce: Simplified Data Processing on Large Clusters' [7]. This was the next stage of progression from traditional analytics methodologies explained in previous sections. Similar principles of Google File System and MapReduce were developed into open source tools Apache Hadoop Distributed File System (HDFS) and Apache MapReduce and they were collectively called Apache Hadoop. Both of these tools were designed to work on commodity hardware and to work in unison on a cluster of machine to provide a distributed filesystem which supported MapReduce principle of breaking the work in smaller pieces to be done in parallel on individual cluster machines (Map) and then join the work together to provide a final result (Reduce). Gradually, lot of other opensource tools were developed to work with HDFS and MapReduce to handle different types and formats of data. Tools like Apache Hive[6], Apache HBase[7] were developed and were widely used for providing a relational access point to structured and non structured data respectively. There are numerous other tools which were developed other than these to provide a wide spectrum of flexibility to Hadoop platform to deal with nearly any type, format or data source. Namely, Apache Pig[8], Apache Flume[9], Apache Kafka[10], Apache Sqoop[11].

[2]IBM Netezza designs and markets high-performance data warehouse appliances and advanced analytics applications

[3]Teradata Corporation is a provider of database-related products and services.

[4]Storage Area Network

[5]Object storage also known as object-based storage is a computer data storage architecture that manages data as objects

[6]Apache Hive is a data warehouse software project built on top of Apache Hadoop

[7]Apache HBase is a data warehouse software project built on top of Apache Hadoop for NoSQL databases

[8]Apache Pig is a high-level platform for creating programs that runs on Apache Hadoop

[9]Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data

[10]Apache Kafka is an open-source stream processing platform developed by the Apache Software Foundation written in Scala and Java

[11]Sqoop is a command-line interface application for transferring data between relational databases and Hadoop

2.4 Challenges with contemporary technologies

Hadoop MapReduce and HDFS were considerably used, with other required Hadooop tools based on need, and are still utilized substantially for a wide variety of big data processing, transformation and analytics. As explained previously Social Media domain is so dynamic and growing that the amount of data being generated [6] grew so large that MapReduce started to appear as it has reached to maximum of performance it can provide and there was a need for an alternative [9]. On the other hand, however, HDFS still remains a very important pillar in this domain. Continuous advancements to increase the performance of HDFS are still being made to increase the I/O performance of the data like storing data in Apache Parquet[12], Apache Avro[13], Apache ORC[14] file formats to serialize the data and to increase the performance while reading bulk of data. In addition to that different file compression formats like normal gzip, snappy etc. are also used these days to compress the files while writing them on HDFS to impart a shorter footprints of file sizes which in turn increases the performance on writing and reading files to and from HDFS [2].

MapReduce has certain challenges when the amount of data grows too big. The fundamental problem with MapReduce is that in principle it creates multiple stages for any type of query of data transformation and all of the data output of these intermediate stages is stored on HDFS and then it is read back from HDFS for subsequent stages. Because, MapReduce works on these files directly from HDFS in principle it spends a lot of time doing I/O on HDFS and eventually the disk. This performance is good for a certain amount of data but as we established that Social Media data is huge and ever growing, MapReduce is not a viable solution because of low performance. There are various use cases on social media data analytics where the results are expected to be retrieved very quickly. For example, There are use cases where a 3D model is generated to visualize the social media data usage and it demands a very high performance throughput from the system on a very huge size of data sets [18]. Many social media data sources are not really static data and are streams of data like Live Chat data or Live Google My Business Reviews where if the analysis is to do live reporting of the data as it is being generated MapReduce may not be the optimal choice.

2.5 Apache Spark and its benefits

Apache Spark[15] was an open source tool developed keeping these shortcomings of MapReduce in mind [9]. Spark on one hand works on the similar concept of MapReduce but the data for different Stages of the execution is not stored on HDFS or actual disks. Spark attempts to store as much data as possible in the Memory of the distributed cluster. Because Memory (RAM) are must more faster than any kinds of disks SATA, SSD etc. the performance of Spark is much faster than MapReduce jobs [9]. Spark necessarily doesn't require a Cluster of machine and can work on single nodes as well. However, the real throughput and performance of Spark data processing, transformation and analysis jobs is when running it on distributed system. Spark provides fundamental data structures like Resilient Distributed DataSets (RDDs), DataFrames and DataSets which can work on highly distributed systems and also provide immense amount of APIs to make the data processing quicker and easier to Develop [5]. Spark doesn't have a specific requirement to be used on a Hadoop Cluster but in the interest of this work we'll focus only on applications of Spark where HDFS provides the distributed file system to work hand in hand with Distributed Data of Spark Data types. In addition to that, if required, Spark can also work with other data types directly like Object Storage, File Data as well. Spark, can also work with Mesos or in standalone mode on Cloud.

There are many other features that make Spark an extremely viable solution for Social Media Analytics. Hadoop, on one hand, resolved a lot of issues with having different formats of data and types of data but there are still a lot of other analysis which require data to be learned on the fly etc. Spark provides a lot of libraries and APIs which can directly handle these different sources of data. Spark Streaming provides APIs to read data from streams of data like Twitter Stream etc, Spark SQL[16] provides APIs to run SQL like queries on data retrieved, Spark Machine Learning[17] library provides APIs to create models on the data to make prediction analysis and finally Spark GraphX[18] library provides APIs for graph data and for graph parallel computation.

3 USE CASE

At this point we have established the limitations of Traditional Analytics technologies and methodologies which are limited to Traditional Data analytics needs, whereas, for the ever growing and extremely dynamic data of Social Media we need much more than Traditional Methodologies. Even the contemporary tools which are widely used in Social Media lack performance and supported features which can fit all kind of data analysis needs of Social Media [13]. Two substantial usages of Social Media data other than many are collecting data to find insights on how the product itself can be improved or to find how the product is doing in the market and to advertise it better.

3.1 Product enhancements

A use case of the first category is reviews. Yelp[19] and Google My Business[20] are crowd sourcing sites which helps getting reviews from all the users of Yelp and Google about various businesses. A substantial part of these businesses are restaurants, where users can provide their feedback of all of these restaurants in form of textual information as reviews. And can also provide ratings in stars to the restaurants. This data has a great potential of providing great insights of what the restaurants can improve upon. We do know that there's a lot of research and technologies available for Natural Language Processing (NLP) and Sentiment analysis. But the problem here is not how to find insights, that is the data science

[12] Apache Parquet is one of many serialization formats in Hadoop
[13] Hadoop serialization format
[14] Apache ORC is serialization format used in Hadoop
[15] Apache Spark is an open-source cluster-computing framework
[16] Apache Spark SQL Library
[17] Apache SPark Machine Learning Library
[18] Apache Spark Graph Library
[19] Yelp is an American multinational corporation headquartered in San Francisco, California
[20] Google Application for businesses

part of the problem. The problem is data engineering and the sheer amount of data that is being generated. With Spark this data can be ingested to high performance clusters directly via Apache Flume and Kafka to Spark Streaming APIs. By applying the Lambda architecture [17] spark can provide a continuous ingestion of data at real time and it can processed, transformed (possibly NLP) and can be aggregated to generate reports in real time for different businesses. This is not possible with any of the traditional analytics technologies or even contemporary Hadoop MapReduce.

3.2 Descision Making for Marketing

Another use case for the second category is marketing. Many Social Media Sites are being used to market products these days in form of advertising. It could be a sponsored post in someone's timeline (Facebook, Instagram) or it could simply be an ad which shows up on ad space on your webpage or in the social media application. This advertising depends highly on conversion rate of any user i.e. the user actually clicks or visits the site or product being advertised. It is highly possible that user might not be interested in that kind of product at all. There are some lower level analytics done in the browsers themselves these days where cache of the browsing history of any user can be utilized to show an ad of a product which the user was looking at sometime back. This particular advertising is called Behavioral Retargeting [20]. But that's very straight forward problem to solve and there are many 3rd pary providers like Adroll, Retargeter who provide these services. The advertising can be improved to a very larger extent if the social media interactions of the users like what kind of video the user liked, what photos user is more interested in, what kind of demographics and geography the user has affinity to [4]. Numerous such statistics, if processed and mined, a good machine learning model can be created using machine libraries of spark to get this data in real time via Spark Streaming APIs and after processing, analyzing data with lambda architecture, final reports can be generated or if required actions can be triggered in real time to choose what category of the ads for a particular user has a high chance of getting a conversion. This again is something where considering the amount of data and the very high throughput expectancy its not possible to achieve this with traditional analytics technologies [13].

4 FUTURE WORK

After this work it can be said with at most ease that Apache Spark is one of the best available technology for Social Media Analytics and as we've have established its viability in some use cases as well, a good meaningful next step on this work would be to implement a spark project on a virtualized environment and integrate it with a Social Media data source. This can help quantify the performance and other aspects of application of Spark in Social Media and Big Data.

5 CONCLUSION

After exploring all types of data available, traditional and contemporary, specifically Social Media, we established the enormity, wide variety and growth rate of Social Media Data. We also examined the shortcomings of the traditional technologies and even the contemporary big data methodologies and how they are not a best fit for the analytical and data processing needs for Social Media data. After looking closely at the wide set of features and custom solutions that Apache Spark can provide we were successfully able to showcase how Spark can be a best bit for all the data processing and analytics needs of Social Media data. We also discussed the application of Spark on Social Media Data with a few example use cases. The use cases we discussed are much broader and are a simple overview of how Spark can be utilized best with the contemporary data analysis needs with the highly volatile and exponentially growing social media data of various types, sources and formats.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his extensive support in this work.

REFERENCES

[1] Abir S. Al-Harrasi and Ali H. Al-Badi. 2014. The Impact Of Social Networking: A Study Of The Influence Of Smartphones On College Students. *Contemporary Issues in Education Research (CIER)* 7, 2 (2014), 129–136. https://doi.org/10.19030/cier.v7i2.8483

[2] Vaddeman B. 2016. *Beginning Apache Pig* (1st. ed.). Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2337-6_15

[3] Bogdan Batrinca and Philip C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *Springer London* 30 (2015), 89fi?!116. https://doi.org/10.1007/s00146-014-0549-4

[4] Ricardo Limongi Frana Coelho, Denise Santos de Oliveira, and Marcos Inacio Severo de Almeida. 2016. Does social media matter for post typology? Impact of post content on Facebook and Instagram metrics. *Online Information Review* 40 (2016), 458–471. https://doi.org/10.1108/OIR-06-2015-0176

[5] Jules Damji. 2016. A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets. (2016). https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html accessed 2017.

[6] Sarah Dawley. 2016. A Long List of Facebook Statisticsfi!?And What They Mean For Your Business. (2016). https://blog.hootsuite.com/facebook-statistics/ accessed 2017.

[7] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[8] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google File System. *SIGOPS Oper. Syst. Rev.* 37, 5 (Oct. 2003), 29–43. https://doi.org/10.1145/1165389.945450

[9] Satish Gopalani and Rohan Arora. 2015. Article: Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means. *International Journal of Computer Applications* 113, 1 (March 2015), 8–11. Full text available.

[10] INTEL. 2012. Cloud Computing Research for IT Strategic Planning. (2012). https://www.intel.com/content/dam/www/public/us/en/documents/reports/next-generation-cloud-networking-storage-peer-research-report.pdf accessed 2017.

[11] George J. Trujillo Jr., Charles Kim, Steven Jones, Rommel Garcia, and Justin Murray. 2015. *Virtualizing Hadoop* (1st. ed.). VMware Press, 800 East 96th Street, Indianapolis, Indiana 46240. http://www.pearsonitcertification.com/articles/article.aspx?p=2427073&seqNum=2

[12] Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53, 1 (2010), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[13] K. Krishnan. 2013. *Data Warehousing in the Age of Big Data* (1st. ed.). Elsevier Science, 225 Wyman Street, Waltham, MA, 02451, USA. https://books.google.com/books?id=8ngws8f_lNsC

[14] Amanda Lenhart. 2015. Teens, Social Media & Technology Overview 2015. (2015). http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/ accessed 2017.

[15] NCSU.EDU. 2014. Social Media Data Research and Use. (2014). https://www.lib.ncsu.edu/social-media-archives-toolkit/research-and-use/research accessed 2017.

[16] Abderrazak Sebaa, Fatima Chikh, Amina Nouicer, and Abdelkamel Tari. 2017. Research in Big Data Warehousing using Hadoop. *Journal of Information Systems Engineering & Management* 2, 10 (2017), 1. https://doi.org/10.20897/jisem.201710

[17] Gwen Shapira. 2014. Building Lambda Architecture with Spark Streaming. (2014). https://blog.cloudera.com/blog/2014/08/building-lambda-architecture-with-spark-streaming/ accessed 2017.

[18] Zachary Weber and Vijay Gadepally. 2014. Using 3D Printing to Visualize Social Media Big Data. *CoRR* abs/1409.7724 (2014), 1. http://arxiv.org/abs/1409.7724

[19] Yusheng Xie, Zhengzhang Chen, Yu Cheng, Kunpeng Zhang, Ankit Agrawal, Wei-Keng Liao, and Alok Choudhary. 2013. Detecting and Tracking Disease Outbreaks by Mining Social Media Data. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, Beijing, China, 2958–2960. http://dl.acm.org/citation.cfm?id=2540128.2540556

[20] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How Much Can Behavioral Targeting Help Online Advertising?. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 261–270. https://doi.org/10.1145/1526709.1526745

Big Data Platforms as a Service

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

ABSTRACT

Big Data platform solutions allow data producers to use data to the fullest potential by combining processing engines with storage solutions and analytic technologies. Pharmaceutical clients are looking into platform solutions to safely store, analyze, and use clinical trial data, experimental data, drug development studies, drug production, regulation, and a number of other outlets. Just a few of the benefits of using a platform solution to manage these data outlets are possibly not having to change current work processes, that management and other research groups can access and use data without needing special access to systems, and scalability of storage and analytic components is seamless. The problems faced to implementing big data platform solutions include the selection of a platform vendor, the design of appropriate data architecture, and establishing effective user interfaces.

KEYWORDS

i523, HID313, Big Data, Platform, Cloud Architecture

1 INTRODUCTION

Most pharmaceutical companies have adopted one or many Laboratory Information Management Systems (LIMS) and/or Electronic Laboratory Notebooks (ELN). These systems are often implemented as standalone systems within a single Research and Development (R&D) group or even within a single laboratory. A problem seen in large- or mid-sized pharmaceutical companies is that different research groups within the same organization often implement isolated LIMS or ELN. This severely restricts data sharing and reuse between groups which leads to many problems such as the same experiment being run multiple times between different groups, regulatory inefficiencies in tracking sample use and storage, and bottle necked development cycles due to missing data.

One of the emerging strategies to combat the problems arising from isolated systems is to combine systems using cloud computing. Platform as a Service (PaaS) provides an environment for the development and execution of applications and software tools. The platform is the heart of a cloud computing infrastructure that enables the software on-top as well as any data created to be accessed and used by a multitude of users [8].

2 IMPORTANCE OF PLATFORMS

Many organizations struggle to share data and processing tools among researchers. PaaS provides a method of better resource utilization while reducing maintenance costs [7]. As pharmaceutical companies collect larger and larger masses of data through LIMS, ELN, and other systems the need for scalable storage becomes inescapable. Cloud storage available with the implementation of a PaaS solves current and predictable future data storage needs as clinical trial data becomes truly digital and full genome analysis becomes more available. The surge of stored data requires access to tools with the capability of pulling insights from the data. These analytic tools are available in familiar formats that statisticians know and love such as SAS, but new analytic tools have been built into platform environments as well as pushed the development of new market players like Tabeau and Spotfire [10].

A pharmaceutical company's R&D group is made up of several diverse units such as analytical chemistry, oncology, genomics, etc. Each of these groups has their own set of unique requirements and thus require multiple solutions to be implemented across the R&D organization. A problem arises now when an FDA regulator enters the lab space and requires an audit trail for a single sample. The sample was aliquoted and distributed across several groups and R&D management needs to be able to prove to the FDA regulator that the sample has been used only for its designated purpose and has been properly destroyed. The sample's use is recorded in several different LIMS and an ELN which the R&D manager does not have access to. With a properly implemented PaaS the manager can print the usage audit trail from each system without accessing them individually. The manager can pull destruction records and storage locations from current inventory and deliver these records to the FDA regulator without directly contacting any of the lab groups.

3 IMPLEMENTING PLATFORMS

Implementing a platform raises a number of concerns around security, selecting the right solution, designing the data architecture and associated relationships, and planning the user interface. All of the large platform providers have invested enormous amounts of resources into assuring the security of their data storage solutions. The right solution might be based on available applications, the storage solution's design, the cost, the learning curve for use, or a number of other client based requirements. Data architecture has the overarching purpose to design the data warehouse solution without limitations to growth, analysis tools, or query speed. User interface depends mostly on the user requirements, it could be driven by how much visibility is needed and how read and write privileges are designated.

The overarching concern with storing data outside of the organization is security. Numerous methods have been developed to assure cloud security such as integrated stacks used by Google and Microsoft Azure and Service Level Agreements (SLAs) [2]. Cloud companies are required to maintain high security at all levels. Google runs various vulnerability reward programs that pay developers, hackers, and security experts for finding security bugs. In addition to the product bugs, Google also maintains high security at their data centers, which includes laser beam intrusion detection,

multi-factor access control, and biometrics to a limited population of less than 1% of Googlers [4].

4 IDENTIFYING THE RIGHT PAAS

Every organization has a unique set of user requirements and every organization shares a certain number of user requirements. Something as simple as requiring a username and password to access content is a requirement shared across the great majority of systems while the need to create complex animal breeding plans that produce offspring with genetic content for 20 specific alleles may be a requirement for one unique client. A market analysis weighing a platform's capabilities against the organization's requirements will often help to narrow down this expanding market. Some of the largest PaaS providers are Microsoft, Amazon, and Google.

Microsoft big data solutions have taken advantage of open source technologies by setting Hadoop as the center of their big data platform. Hadoop is implemented through Hortonworks Data Platform (HDP) which has been developed as an open source solution with Apache and other open source components. Microsoft allows cloud and on-premise implementation, but generally local environments are only used as proof of concept testing. Microsoft platform solutions allow for data to be manipulated and used in Microsoft tools such as Sharepoint and Excel while big data analysis, visualization, and mining can be performed using SQL Server Analysis Services or HDInsight. The Hadoop-based platform has no limitations with structured or unstructured data, a number of additional tools are available for data storage, and efficient queries provide a potential boost to discovery. Microsoft Azure storage runs $40 a month per 1TB and employs a pay for use plan to resource use within the platform's toolbox [6].

Amazon Web Services (AWS) offers data storage solutions in NoSQL and Relational Database models. Interactions with these data engines can be done using Hadoop, Interactive Query Service, or Elasticsearch. Amazon has designed their storage sources in such a way that clients can use any preferred open source application, but Amazon has also developed a toolbox of analytic tools. Amazon offers data warehousing through Amazon Redshift, which allows for management, query, and analysis at the petabyte-scale. Amazon storage runs around $80 a month per 1TB. AWS offers Business Intelligence, Artificial Intelligence, Machine Learning, Internet of Things, Serverless Computing, and a number of data interface tools available in a pay-as-you-use billing form [1].

Google Cloud Platform (GCP) offers a complete end-to-end data storage solution, which allows the use of GCP developed systems and open source tools. BigQuery is Google's data warehouse tool which is serverless and requires no infrastructure management with the assist of Google Cloud Dataflow. Dataflow eliminates the need for resource management and performance optimization. GCP storage runs $10 a month per 1TB. GCP has a number of applications for data manipulation. Dataproc allows dataset management through Hadoop and Spark, data visualization can be generated through Datalab, Data Studio, and Dataprep which are all Google developed applications [3].

5 DESIGNING THE DATA ARCHITECTURE

All data storage solutions from relational databases to noSQL data stores to cloud data warehouses have to start with a defined architecture. The data architecture model will illustrate how data components will be organized and connected. The mindset of a data architect should be focused on reducing the complexity of a data model while maintaining the highest level on utilization. This can be a fine line to walk as a designer. Complexity can be reduced by breaking user requirements down to the most basic and generalized principles to define the simplest data modules. An example of this might be a system that requires a number of different requests and instead of designing a component for vendor requests, user requests, and management requests the component is designed for request and request type. This generality allows for easy future scaling or additional system requirements not yet defined. Cloud systems maintain high utilization by manipulating data using strategic layering. One layer for storage, one layer for defining storage keys, another for combining query tools, another for consolidating query results and so on. With the more established cloud offerings a lot of these layers have already been supplied, but the transitions and interconnections still have to be outlined by a designer [9].

6 DESIGNING THE USER INTERFACE

A system's user interface (UI) must be laid out in a simple and intuitive manner that allows users to perform the tasks required while exploring new insights provided by generated data. There are a number of influences leading to the development of user interfaces such as familiarity; users are familiar and comfortable performing a search in Google or Amazon interfaces and maintain the same high expectation with their working environment. If a user requires sample tracking or auditing, they may relate the need to how a package is tracked with FedEx or UPS and expect the same level of access and insight to sample tracking within their working environment. Users may even have an information management system that they use and are comfortable with so switching to a new UI can be daunting as it requires additional training and most likely new work processes.

UI developers have the challenging job of creating the face of an application. A poorly designed face may not attract as many customers as something with a higher graphical output. Even a strong performing system can be downgraded or completely ignored by users if its front end is poorly laid out. Considerations for a UI design include font-size, space between elements, interactive space, and line-width which can all differ across devices such as between a tablet, desktop, or smartphone [5].

7 CONCLUSION

As more and more companies realize the value of their data, platforms and associated tools become more and more vital to organizational success. The pharmaceutical industry knows that data is king, but is experiencing major bottlenecks in deploying platform solutions for the reasons discussed: the cost and complexity of implementation, the concern over security, the frustration of changing or creating new work processes. Current information management systems help scientists and researchers work exponentially faster than they ever could on paper, but current systems are not designed

to facilitate sharing of ideas. This is where platforms come in. A regulatory supervisor should not need training on every information management system to effectively regulate the use and disposal of clinical samples. A laboratory technician should not need to wait for specific system privileges to access a study that the organization did in a different lab space, whether it's in the same building or on the other side of the globe. Platform services are allowing scientists and managers to share ideas more efficiently than they ever have before and the pharmaceutical industry has the potential to exploit this new technology to improve life expectancy, make drugs safer, and research smarter.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor Von Laszewski and Teaching Assistants Hyungro Lee, Juliette Zerick, Saber Sheybani Moghadam, and Miao Jiang.

REFERENCES

[1] Amazon. 2017. Big Data on AWS. Website. (Oct. 2017). https://aws.amazon.com/big-data/

[2] Valentina Casola, Alessandra De Benedictis, Massimiliano Rak, and Villano Umberto. 2014. Preliminary design of a platform-as-a-service to provide security in cloud. *ResearchGate* 2014, 1 (01 2014), 752–757. https://www.researchgate.net/publication/289573602

[3] Google. 2017. Big Data Solutions. Website. (Oct. 2017). https://cloud.google.com/products/big-data/

[4] Google. 2017. Google Security Whitepaper. Website. (Oct. 2017). https://cloud.google.com/security/whitepaper#state-of-the-art_data_centers

[5] Miroslav Macik, Tomas Cerny, and Pavel Slavik. 2014. Context-sensitive, cross-platform user interface generation. *Journal on Multimodal User Interfaces* 8, 2 (01 Jun 2014), 217–229. https://doi.org/10.1007/s12193-013-0141-0

[6] Microsoft. 2017. Understanding Microsoft big data solutions. Website. (Oct. 2017). https://msdn.microsoft.com/en-us/library/dn749804.aspx

[7] Sungyoung Oh, Jieun Cha, Myungkyu Ji, Hyekyung Kang, Seok Kim, Eunyoung Heo, Jong Soo Han, Hyunggoo Kang, Hoseok Chae, Hee Hwang, and Sooyoung Yoo. 2015. Architecture Design of Healthcare Software-as-a-Service Platform for Cloud-Based Clinical Decision Support Service. *Healthcare Informatics Research* 21, 2 (April 2015), 102–110. https://doi.org/10.4258/hir.2015.21.2.102

[8] Arto Ojala and Nina Helander. 2014. Value creation and evolution of a value network: A longitudinal case study on a Platform-as-a-Service provider. In *47th Hawaii International Conference on System Science*, Vol. 47. 47th Hawaii International Conference on System Science, Waikoloa Village, HI, 975–984.

[9] Jerome H. Saltzer and M. Frans Kaashoek. 2009. *Principles of Computer System Design: An Introduction.* Morgan Kaufmann, Burlington, Massachusetts. https://doi.org/10.1016/B978-0-12-374957-4.00010-4

[10] Domenico Talia. 2013. Clouds for scalable big data analytics. *Computer* 46, 5 (2013), 98–101.

Roles and Impact on Mobility Network Traffic in Big Data

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

ABSTRACT

Mobility network traffic is both large in quantity and diverse in data types. Big Data has an opportunity to align these data types in such a way to provide meaningful information to the network provider. In short, there is an opportunity for Big Data to impart wisdom.

KEYWORDS

i523, hib315, Big Data, Mobility Network Traffic, Network Forecasting, Data Types

1 INTRODUCTION

At the core of Big Data is a challenge. A challenge of exploration "of the complexities inherently trapped in data, business, and problem-solving systems". [1] Which is by definition, "Big Data".

Imagine a world where your business decisions relate to data sources that range from a flat file from a third-party vendor to millions of internal data records every day, nearly every hour. Add to this, some data sources might "round up" the data, while others relate the data (traffic) to a different geographic standard then others. So it is in the world of mobility network traffic.

2 DATA TYPES AND CHALLENGES

Mobility network traffic providers generate CDR (Call Detail Records) every time a device establishes a connection. These CDRs that are produced by the network equipment provide details about the connection - cell site locations, length of call and device information, including the duration of the call along with other information. It is from these records that the network providers gather, *clean* if need be, consolidate and extrapolate the needed information to bill the customer.

In terms of the CDR data, a large telecommunication provider will create millions of these records every day, even every hour. For companies that have over 50 million devices to manage, and each device can create dozens of records each day, the numbers of records and the size of the data is tremendous. However, with all this data, the management of the records by sheer quantity can lead to qualitative challenges. For example, by the time the millions of records are consolidated to generate files that are more manageable, data details can be lost. While CDRs tell us a great deal, there is much that they do not tell a provider. Therefore, other data sources are used, like data from the network, which provides precise traffic metrics.

This additional network data does not come from the creation of CDRs but is rather collected from the numerous sectors within a mobility network. A sector is a collection of cellular towers and these sectors are in turn gathered together and feed metrics into what are referred to as data collectors. Therefore, the collectors are related to the network vendors that build the equipment. As a result, if the network has more than one equipment vendor, a challenge is to make sure the vendors measure or collect traffic consistently across the network. Once we are insured of consistent measurement of the data from the collectors, we can then consistently map the data into agreed upon geographical areas, known as sub-markets or markets.

This additional network data is free of the challenges and limitations of the CDR based data, this data however, only provides simple traffic measurements. For example, we now know the voice traffic measured in minutes, or the megabyte (MB) traffic in California or South Dakota. But it doesn't tell us the device type or any customer specific data like the CDR data does.

Adding to the challenge, companies like Verizon and AT&T are changing to unlimited plans - which allows the customer complete data freedom, offering package deals with video services and even offering free traffic based on cell phone apps (HBO for free on your device) - So gathering meaningful data on this type of traffic, requires a data set that is much different than simply looking at network or CDR traffic. That is, we need to look at the bits and bytes. We need a much deeper dive into the traffic to start to pull more specific information. For example, we can look at the data packet headers and leveraging an involved process can start to glean an understanding about the network traffic that provides us details and specifics around this big data. For example, we can get data regarding how much traffic is video (directly in relationship to promotions like free HBO), or how much traffic was browsing the web, instant messaging, photo files, VoIP (Voice of IP) and many others.

Additionally, the customer landscape has changed which makes traditional analysis more challenging. For example, in years past, most of the mobility subscribers were *post-paid*. That is, they paid after the actually activity took place. Most mobility subscribers used their mobile device last month and then received their bill this month. Today we have pre-paid customers, wholesale customers and even customers that simply monitor their packages, dog-collars, vending machines and track delivery trucks. We call this the Internet-Of-Things (IoT).

With IoT, lots of projections abound and here is one: "roughly 23 billion active IoT devices by the year 2019 and spending on enterprise IoT products and services will reach $255 billion globally by 2019, up from $46.2 billion this year." [3]

Also network providers have learned that nothing puts more traffic on the network like video. Video based apps, like Facebook and You Tube directly impact network traffic. [2]

The impact of apps on the mobility network is significant with no end in site: "when it comes to reaching consumers in mass, the market has confirmed what we have known all along - that we are all building and investing into a platform that can reach heights we

may have never seen before. That, to me, is "The WhatsApp Effect", and there is no turning back now." [4]

As shown in Figure 1. You can see the projected video usage increase, by a percentage of the total network traffic over the next five years.

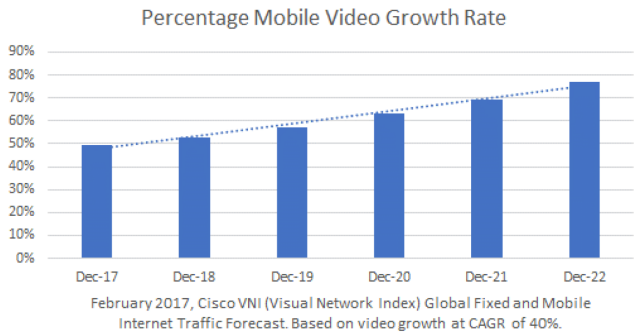

Figure 1: Cisco

This leaves us with yet another type of data impacting the mobility network that is neither network traffic or data around the traffic types. This is data from the applications. Most of the data from a mobile device is tied to one of many "apps". While the process is wrought with challenges, larger network providers will invest in diving yet deeper into the network and traffic types in order to have a better understanding around traffic specific to apps. Prudence would dictate to a network provider that it is best to know what is on their network. However, strict legal and customer privacy laws, along with application vendors working independently thus not coordinating with network providers, leads to numerous data challenges. While far from a perfect process, gathering as much app level data is critical to the management of any mobility network. Not only to the management of the mobility network but it provides value added information that could impact marketing plans, finance and organizations like strategic planning and technology planning. All in an effort to understand and provide outstanding customer service.

In addition to the efforts of the mobility network provider in gathering data around apps, there are other related data options. Like the saying, "there is an app for that", there are other means of gathering such data. App Annie is one of many companies that provide data related to apps. Both in terms of the numbers of uploads of an app as well as gathering high level app related metrics. As an app developer, imagine knowing the amount of uploads of your app, geographical upload metrics as well as revenue related to the uploads.

Similarly other companies gather app related traffic as a result of their own app that manages customers data packages so that the customers do not use too much data. There are also companies that inform the customer of their intention to gather data based on their usage. This is usually done in such a way that there is no one customers' data that is identified but data from many customers that is combined to provided analytics. Still others have apps that manage the efficiency in not draining too much of the device's battery.

These app level data sources can be critical when a network provider is trying to identify traffic that is no longer on the mobility (cellular) network but has moved to Wi-Fi. Once the traffic is off the mobility network you no longer have data regarding it. All the traditional network data sources are of little, to no, benefit. This importance is magnified when looking at particular apps that can add significant data traffic to the mobility network. For example, Netflix is a heavy Wi-Fi leveraged app but imagine if a percentage of the traffic rolled to the mobility network. So keeping a close eye on the traditional video streaming apps, and it's percentage of usage on Wi-Fi, is a wise decision. As a result, the additional sources can prove critical in building knowledge around your data.

3 CHALLENGE AND CONCLUSION

For mobility network providers, what is the Big Data challenge here? What is the missing piece to the providers that Big Data has an opportunity to help with, if not answer? Providers already have access to network traffic data, along with data around traffic types which is above the OSI Model Network Layer (Open Systems Interconnection) to provide some insights into traffic types; web browsing traffic, VoIP, video, and even some data around traffic related to apps. The challenge for Big Data is to take all of this data and give network providers accurate analytics on - *customer behavior*!

Can it be done? I believe, with the use of data holistically and with data-driven discovery, it can. However, it is important to note that in order for this to be successful, you have to have a solid understanding of the data itself. It requires an intimate knowledge of the data, the sources, and any underlying limitations and collection challenges. Additionally, it is critical to have substantive data storage capabilities, like data lakes.

A holistic view of the data is to include all the data sources; network data, traffic type data, app level data interrelated and connected hierarchically, so that when you see a jump in the network traffic, you trace the traffic type and app level, which can then lead to accurate deductions to explain the, aberration, one such as, *The Ice Bucket Challenge*, an innocuous social experiment played out on Facebook that demanded a tremendous amount of network capacity. This comprehensive, holistic approach is the only way to paint an accurate picture of user behavior, taming "Big Data" into a beast that can be interpreted. And as a result, helping understand - customer behavior.

At this point we have built a relationship between the various data sources and have let the data drive the results. It's from this process in which we have gained an important business benefit - wisdom. Wisdom gained from a data-driven discovery that can be applied directly to the mobility network itself. From a Big Data challenge, and given data knowledge, we aligned the data and let the data "tell" us the impacts on the network. This wisdom provides us with one last critical benefit for any mobility network provider - a better bottom line, which as they say, is the bottom line.

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] Longbing Cao. 2017. Data Science: Challenges and Directions. Magazine. (Aug. 2017). https://doi.org/CommunicationsoftheACM

[2] Cisco. 2017. *Cisco - Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper.* techreport. Manufacture. https://doi.org/Manufacture

[3] Jay Schofield. 2015. Big Data Challenges Wireless Networks, CIOs. Webpage. (Feb. 2015). http://www.systemid.com/learn/big-data-challenges-wireless-networks-cios/

[4] Semil Shah. 2014. The WhatsApp Effect. Webpage. (March 2014). https://doi.org/TechCrunch

NoSQL Databases in Support of Big Data and Analytics

Uma M Kugan
Indiana University
711 N Park Avue
Bloomington, IN 47408, USA
umakugan@iu.edu

ABSTRACT

The data volume is increasing at high velocity and it comes from various sources with different formats. These data no longer fits into defined structure and hence the need for handling the big data using NoSQL. This paper will highlight on what is NoSQL and where and when it should be used for and also why Big Data can not be handled in traditional RDBMS.

KEYWORDS

i523, hid323, NoSQL, Bigdata, RDBMS

1 INTRODUCTION

RDBMS have always been the preferred method of storage for many years and its powerful Query language made it very user friendly. Data has grown exponentially in a past decade due to the growth of social media, e-commerce and web applications which posed a big challenge for the traditional databases. Need of the hour is not just to limit the data within the structure, but also ability and flexibility to read and store data from all sources and types, with or without structure. Companies that has larger amount of unstructured data are shifting away from traditional relational databases to NoSQL [5]. There are lot of limiting factors in these databases for Big Data especially Structured schema which was one of the main reason for RDBMS to scale it for larger databases [10].

2 LIMITATION OF RDBMS

Choice of database chosen depends on their data model, data access and data latency. But in this era, every organization needs all three at the same time and which can not be provided by traditional databases [11].

Scalability - RDBMS are designed for scaling up meaning if storage needs to be increased, we need to upgrade other resources in the existing machine whereas in NoSQL we just have add additional nodes in the existing cluster.

Acid Compliance - RDBMS are always acid compliant and which of course is its strength to process transactional data while the drawback is it can not handle larger volume of data without impacting the performance. If there are use cases where we do not require ACID compliance and where it has to handle huge volume of data in significantly very less time, then NoSQL is the solution.

Complexity - RDBMS stores the data in defined, structured schema in tables and columns. If the data can not be converted to store in tables, it becomes cumbersome to handle such situations.

Time Consuming - Analyzing data in real time is highly impossible with RDBMS and no one have time to wait for longer load schedules in traditional way of data warehouse and ETL.

3 NOSQL

"The term NoSQL was first used by Carlo Strozzi to name a database management system (DBMS) he developed. This system explicitly avoided SQL as querying language, while it was still based on a relational model" [1]. The term NoSQL means that the database does not follow the relational model espoused by E.F Codd in his 1970 paper, "A Relational Model of Data for Large Shared Data Banks which would become the basis for all modern RDBMS" [9]. NoSQL does not mean NO to SQL. It means Not Only SQL. NoSQL means storage is just nonvolatile object store with no maintenance concerns. Most NoSQL Databases are open source which allows everyone to evaluate the tool of their choice at low cost. NoSQL databases, because of it's simpler data model, it does not need DBA's to maintain the health of the database. NoSQL databases are widely used in big data and in real-time applications.

4 NOSQL TYPES

In Edlich et al. identify four classes of NoSQL systems as 'Core-NoSQL' systems. NoSQL systems are primarily differentiated by data model and also on how the data is stored. They are Key-Value stores, Wide column stores, Graph databases and Document stores [3].

Key-Value Stores - Key is the unique identifier or label of an item whose data or its location is stored in the value. It is very basic non relational data types which is most commonly used. Example include Redis, Amazon DynamoDB and Oracle NoSQL.

Wide Column Stores - Every record in the stores may differ in the number of columns. This is very important factor for analytic because it needs very low I/O and also reduces the volume of data that are read to the disk. It is also known as tabular NoSQL database.Examples include HBase, Google BigTable and Cassandra.

Graph Database - As the name indicates, it uses graph structures nodes and edges to represent the data. This is very useful in depicting social relationship, network topology. Examples include Neo4J and DataStax Enterprise Graph.

Document Stores - It stores the data as document typically in JASON or XML format. It is very flexibile and one can easily access the data. This is used widely in many places. Examples include MongoDB and CouchDB.

5 ADVANTAGES OF NOSQL

NoSQL databases differ from traditional databases in features and functionality. There is no common query language, high I/O performance, horizontal scalability and do not enforce schema. RDBMS

scales up vertically making single CPU works faster and performance can be increased adding extra CPU or RAM, whereas a NoSQL database scales horizontally by making many CPUs works together and also by dividing the jobs into multiple chunks [2]. It is very flexible and let the users to decide to use the data the way they want. Data localization in NoSQL databases is acheived by distributing it across many geographic regions. NoSQL databases does not have need specific applications or hardwares to implement replication [6]. Since NoSQL does not enforce atomicity and hence it is not reliable where data accuracy is very critical. The main advantage of NoSQL is its data is always replicated on each node and so the data is always available and there is zero downtime. RDBMS supports master-slave architecture so data can always be written to master and read only data is available in all slave machines whereas in NoSQL, both read and writes are enabled in all nodes. In general most of the NoSQL databases performance is better than SQL databases.

6 NOSQL CHALLENGES

NoSQL databases have created lot of interests in each organization to move away slowly from traditional databases but there are many challenges to overcome. RDBMS are much more matured, been around for many years and the best technical support is available. So there is always fear of unknown until the technology gets widely accepted and used [5]. Most of the NoSQL databases are open source and support and reassurance that any organization gets from their traditional RDBMS vendors are challenged. Even though NoSQL goal is to provide no admin solution,in current trend, it requires lots of skills to maintain and learn. It is highly tempting for any organization to adopt living edge technology, but that adoption needs to be embraced with selection of best tool and with extreme caution [4]. Ad-hoc query analysis is quite complex in NoSQL databases and it requires expertise to write even a simple query.

7 NOSQL FOR BIG DATA

When to choose NoSQL over an RDBMS depends on ACID (Atomicity, Consistency, Isolation, Durability) vs BASE (Basically Available, Soft state, Eventual consistency) and also on the type of the data that the organization is dealing with. Based on the project requirements, If the real time updates is needed to perform data analytics, NoSQL is the solution for applications that receives large volume of data in a real time and where data insights are generated using real time data that was fed [2]. NoSQL is the best fit where the enterprise does not require complex messaging features for publishing/subscribing. NoSQL comes handy where data structure is not restricted by schema(schema less design). Many NoSQL database compromises consistency over availability and data partition [12].

8 HOW TO HANDLE RELATIONAL DATA IN NOSQL

NoSQL database in general can not perform joins between data structures and hence the schema has to be designed in such a way so that it can support joins [11]. Below are the key things that needs to be considered to handle relational data in a NoSQL.

Avoid Sub Queries : Instead of using complex sub queries or nested joins to retrieve the data, break into multiple queries. NoSQL performances are very high when compared to traditional RDBMS Queries.

Denormalize the Data : For faster retrieval of data, it is essential to compromise on denormalizing the data rather than storing only foreign keys.

9 RDBMS TO NOSQL MIGRATION

Database Migrations are always cumbersome and it is better to plan well ahead and take an iterative approach. Based on the need of application, one have to choose which NoSQL database we are going to migrate to [8].

9.1 Planning

The goal of any migration should be better performance at the reduced cost with the newest technology. While migrating from RDBMS, we have to consider volume and source of data that is going to be migrated to NoSQL. All the details should be documented well so that we do not have to face unplanned surprises at the end [7].

9.2 Data Analysis

This is very critical and will help in understanding the nature of the data and how that data is accessed within the application. Based on the analysis of data usage, we will be able to define how data will be read/written which will help us in building a better data model [7].

9.3 Data Modeling

When migrating from any RDBMS, depending on the need of application, we may have to sometimes denormalize the data. In this phase, based on the data analysis and the tech-stream, we have to define keys and values [8].

9.4 Testing

Testing is always very critical and crucial for any migration projects. We have to define all possible test cases and different types of testing: unit, functional, load, integration, user acceptance and smoke testing have to be performed and outputs have to be clearly documented [8].

9.5 Data Migration

Once all the above steps are successfully tested and implemented, next final act is to migrate all data from RDBMS to NoSQL. Post implementation validation has to be carried out to make sure everything went well as per the plan and it has to be monitored for few days until the process is stabilized. If there are any issues with the migration, rollback to original state and root cause analysis have to be performed to identify and fix the issue. Once issue has been fixed, data migration has to be scheduled and this step goes in cyclic unless migration was completely successful.

10 CONCLUSION

With the explosion of the data in the recent years, have paved the big way for the growth of Big Data and everyone wants to move their applications and data into Big Data. Building a big data

environment is relatively very cheap when compared to migrating the existing data in RDBMS to NoSQL. We have to carefully weigh in, understand the data and how the data will be used in the use case to enjoy the full benefit of migrating into No SQL.

ACKNOWLEDGMENTS

My sincere thanks to my mentor and leader Vishal Baijal and to my colleague Michael Macal for their support and suggestions to write this paper and also to my fellow classmate Andres Castro Benavides for his support. My special thanks to Dr. Gregor von Laszewski for his support in fine tuning the paper.

REFERENCES

[1] P. BAXENDALE. 1970. (June 1970). http:/www.seas.upenn.edu/~zives/03f/cis550/codd.pdf

[2] The Enlightened DBA. 2016. *DBA s guide to NoSQL.* Technical Report. DataStax Enterprise, Santa Clara, California. http://www.datastax.com/wp-content/uploads/resources/whitepaper/DataStax-DBAs-Guide-to-NoSQL.pdf?2

[3] S. Edlich, A. Friedland, J. Hampe, and B.Brauer. 2010. *NoSQL: Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Hanser Fachbuchverlag.* Carl Hanser Verlag GmbH —& CO. KG Munich, Detsche.

[4] Rakesh Kumar, Shilp Charu, and Somya Bansal. 2015. Effective Way to Handling Big Data Problems using NoSQL Database (MongoDB).. In *Journal of Advanced Database Management Systems*, Vol. 2. STM Journals, India, 42–48. Issue 2. "https://www.researchgate.net/publication/280622043_Effective_Way_to_Handling_Big_Data_Problems_using_NoSQL_Database_MongoDB"

[5] Neal Leavitt. 2010. Will NoSQL Databases Live Up to Their Promise?. *Computer* 43 (Feb 2010), 12–14. Issue 2. https://doi.org/10.1109/MC.2010.58

[6] MongoDB. 2016. *Top 5 Considerations When Evaluating NoSQL Databases.* Technical Report. MongoDB. https://webassets.mongodb.com/_com_assets/collateral/10gen_Top_5_NoSQL_Considerations.pdf?_ga=2.80016725.353696228.1508503061-18692132.1506430675

[7] MongoDB. 2017. *RDBMS to MongoDB Migration Guide.* Technical Report. MongoDB. https://webassets.mongodb.com/_com_assets/collateral/RDBMStoMongoDBMigration.pdf?_ga=2.27464444.1571008351.1511010273-18692132.1506430675

[8] Nathaniel Slater. 2015. *Best Practices for Migrating from RDBMS to Amazon DynamoDB- Leverage the Power of NoSQL for Suitable Workloads.* Technical Report. Amazon Web Services. https://d0.awsstatic.com/whitepapers/migration-best-practices-rdbms-to-dynamodb.pdf

[9] Carlo Strozzi. 2013. NoSQL–A relational database management system. 2007–2010. (Sep 2013).

[10] Aspire System. 2014. *BigData with NoSQL.* Technical Report. Aspire System, Oak Brook, Illinois. http:/www.aspiresys.com/WhitePapers/BigData_with_NoSQL_Whitepaper.pdf?pdf=nosql-whitepaper

[11] Gaurav Vaish. 2013. *Getting started with NoSQL.* Packt Publishing Ltd, BIRMINGHAM - MUMBAI.

[12] VoltDB. 2017. *SQL vs NoSQL vs NewSQL - VoltDB.* Technical Report. VoltDB. https://www.voltdb.com/wp-content/uploads/2017/05/VoltDB_SQL-vs-NoSQL-vs-NewSQL.pdf

Amazon Web Services in Support of Big Data and Analytics

Peter Russell
Indiana University
petrusse@iu.edu

ABSTRACT

Executives are constantly looking for ways to find the pulse of their competitive landscape along with ways to gauge the sentiment among their customers. The emergence of the Big Data movement has given businesses the unique opportunity to gain perspective on these fronts, in addition to many others. Amazon Web Services has placed itself at the epicenter of this data movement and now offers tools that allows decision makers to quantify their businesses in ways that were previously computationally impossible or were prohibitively expensive. As a result, with Amazon Web Services, companies now have the ability to gain deep insights into customer activity, which can be used as real-time feedback or guidance to make future experiences more personalized.

KEYWORDS

i523, HID334, Cloud Computing, AWS, Big Data Analytics

1 INTRODUCTION

Amazon Web Services (AWS), the cloud service arm of Amazon, is currently the most dominant company in the cloud computing marketplace. With a market share of 31%, AWS holds a larger share than the next three closest competitors (Google, Microsoft and IBM) and contributes $10 billion a year to Amazon[16]. Aside from its financial importance to Amazon though, AWS has become critical for businesses that are looking to gain insights from the data they have at their disposal, especially as this data becomes more abundant [15].

With this business need in mind, AWS offers several products under their "Analytics" platform of services. This is just one of their 18 categories or platforms used to classify their 108 different products. This platform is a particularly interesting area because it is allowing companies to perceive their competitive landscape through an analytical lenses on a scale and frequency not previously seen. Namely, vast data sets in real-time if desired [24].

Our particular focus will be on a high-level description of the products offered in this "Analytics" category, their current utilization by businesses, recent developments in this platform and how it impacts Big Data.

2 ANALYTICAL PRODUCTS

To discuss the impact AWS is having on modern businesses, it's necessary to give a concise description of each analytical service offered. Subsequent sections will then be able to mention these services by name with a basic understanding of that service's function.

2.1 Amazon Athena

Amazon Athena that allows users to analyze data in Amazon Simple Storage Service (S3) as an SQL query. S3 is Amazon's web interfaced data storage and retrieval service, which can be accessed from anywhere, and can be more broadly be described as an Infrastructure as a Service (IaaS). This was designed for queries that may be unique and one-off. Athena remains one of the newest products introduced on the Analytics platform as it was released in late 2016 [1].

2.2 Amazon Elasticsearch Service

Amazon Elasticsearch Service (ES) is a managed service that implements Elasticsearch, which is an open source engine that allows for the indexing of large data sets. This indexing allows for analysis to better understand the events generating the data, such as with a user of an application [3].

2.3 Amazon Elastic MapReduce

Amazon Elastic MapReduce (EMR) is aimed at analysis of large data sets as it allows users to take advantage of a managed Hadoop framework without the traditional setup costs. Hadoop is advantageous over traditional database models because it parses the large data sets over several nodes, allowing for parallel computing and greatly increased efficiency. EMR allows for the iteration over a massive amount of text files while ES is concerned with indexing these files[4].

2.4 Amazon Quicksight

Amazon Quicksight is the data visualization tool that allows for seamless charting and integrating with AWS databases. It also recognizes data types and suggests the best type of visualization for a given analysis [6].

2.5 AWS CloudSearch

AWS CloudSearch is a managed search engine service that can be integrated into an application for a company's users. This allows an easier experience for the user without the company having to dedicate the resource costs that historically came with developing and maintaining the search feature [2]. In fact, AWS CloudSearch uses the same logic and intelligence for search queries that is used on Amazon.com. As one might suspect, AWS CloudSearch is similar to Amazon Elasticsearch Service. However, AWS CloudSearch is fully managed while Amazon Elasticsearch remains the more flexible and popular of the two.

2.6 AWS Data Pipeline

AWS Data Pipeline is designed to ease the maintenance of regular data sets by allowing users to schedule or automate changes to files along with the movement of that data set to other AWS services [8].

2.7 AWS Glue

Broadly speaking, AWS Glue is similar to AWS Data Pipeline in terms of automated transfer and modification of data. However, AWS Glue automates much of this data transformation whereas AWS Data Pipeline offers more flexibility for those who desire it [9].

2.8 AWS Kinesis

The work of AWS Kinesis is likely the most known product of AWS to the common consumer as it is responsible for the processing of real-time data for analysis or alert triggering. A dashboard that displays trending topics on social media or fraud detection at a bank is likely fed by an AWS Kinesis setup [5].

2.9 AWS Redshift

AWS Redshift was created to meet the database storage and maintenance needs of businesses. With Redshift, companies are able to reduce their capital expenditure and time to implementation, both of which could especially critical for nascent companies [7]. This line of business should prove to be increasingly important as data collection by businesses continues to grow. In 2012, it was already estimated that the cost of storage on AWS Redshift was just 10% the cost of traditional database costs [19].

3 RELEVANCE TO BIG DATA: USE CASES

Amazon has stated that they currently have one million active users, which is defined as using their services at least once a month [13]. In exploring current uses it becomes clear that the users are rarely consumers of just one product, opting instead to take advantage of the AWS ecosystem through multiple services. This section will touch upon the most popular AWS products and their interesting uses in the business environment.

3.1 Yelp

Yelp is a search based website that allows users to find different types of businesses while also showing user contributed reviews for these businesses. Started in 2004, Yelp's website now averages 28 million unique mobile users and 83 million unique desktop users per month. These users have contributed 135 million reviews in aggregate [26].

The impact of AWS on Yelp's business planning came when the company was trying to decide how to optimize its advertising revenue [11]. Specifically, Yelp stores log data daily on attributes, such as user location, user query, user clicks and displayed ads. This is all in an effort to better formulate search results given the available data and display ads that are most relevant to users [23].

Of the services discussed earlier, Yelp adopted AWS EMR and AWS Redshift to meet its analytical needs. EMR was implemented to allow multiple teams to analyze the data simultaneously and Redshift was used for easy retrieval. EMR is also used to enhance the user's search experience by returning useful results in the case of misspellings, auto-completion or features such as "People Who Viewed This Also Viewed."[21] As stated earlier, EMR allows this retrieval of information from the stored in nearly real-time. In all, the utilization of these services allows Yelp to be more dynamic as its data analysis time is dramatically reduced while also improving the customer experience and ultimately, retention [14].

3.2 Zillow

Zillow is an online real estate listing marketplace where users can find homes for sale, recently sold homes or foreclosures. One of the largest draws to the site though, is the modeling of a specific property value through a feature they refer to as a "Zestimate." Through the use of AWS Kinesis for data collection and AWS EMR for data processing, Zillow is able to generate home value estimations in virtual real-time for 100 million properties across the United States, which is said to be a function of over 100 input variables[12]. Some of these inputs need to be as real-time as possible, such as recent sales data, for the most accurate estimate, which made Kinesis so impactful [18]. This integration of technologies has dramatically improved their calculation time for these estimates from hours to seconds[12]. Once again, this enhanced user experience through the utilization of Big Data analytics keeps the website relevant and best suited to meet customer needs.

3.3 Netflix

Netflix is a worldwide media provider, offering on-demand movies and shows along with a DVD rental service. Currently, the company has nearly $9 billion in annual revenue with 104 million subscribers [20]. Incredibly, users in aggregate are watching one billion hours of content *a week* and during peak times, Netflix can be servicing over ten thousand streams a second [22]. Perhaps as impressive as the company's success with its user base is the foresight the company had in early as 2008 to begin moving operations to AWS as it began rolling out its internet streaming services. By 2016, they moved their entire infrastructure to the cloud and can have up to a hundred thousand AWS instances running during peak hours [17].

As likely one of the largest AWS users by market capitalization, Netflix casts a wide net across the use of AWS services. By their own admission, insights gleaned from the data they collect play a pivotal role on business and product decisions. Through the AWS Elasticsearch Service, Netflix is able to properly classify its 1.3 PB of data per day (24 GB per second) across different indices, such as viewing activities, error logs and diagnostics [25]. Similarly, Netflix uses AWS Kinesis as the pipeline used to stream this log data and the real-time functionality allows them to identify potential issues immediately [10]. Whether for business or troubleshooting purposes, this data on AWS can be easily visualized through AWS Quicksight for inferences.

Netflix is perhaps the best example of how a company can leverage AWS to outsource the burdens of data management as the volume of data grows. This allows them to focus on the core competencies and customer experience, which like the other examples, maintains or advances their position in the marketplace.

4 RECENT ADVANCEMENTS IN AWS

The most recent advancements in AWS as it relates to Analytics platform have come directly from the introduction of Athena in 2016 and Glue in 2017. Indirectly, AWS has been developing a new product line that is complementary to the Analytics category. In 2016, AWS launched its "Artificial Intelligence" platform, which

is now comprised of seven new services and is clearly an area of growth and focus for Amazon.

Of these new services, Amazon Machine Learning will likely be the most attractive new offering for businesses. This service will allow business users to discover underlying trends in their data and formulate more accurate forecasts.

5 CONCLUSIONS

In these use cases, we've seen that AWS has had a positive impact on Big Data for two reasons. First, businesses are better able to embrace the Big Data movement by making data collection and analysis a priority without the major cost that has historically been associated with such an initiative. Second, we would expect that the successful implementation of cloud analytics will help businesses be more successful, in turn incentivizing them to collect more data and therefore, further expanding the Big Data universe.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the Assistant Instructors for their feedback and help.

REFERENCES

[1] Amazon. 2017. Amazon Athena. Website. (2017). https://aws.amazon.com/athena/
[2] Amazon. 2017. Amazon CloudSearch. (2017). https://aws.amazon.com/cloudsearch/
[3] Amazon. 2017. Amazon Elasticsearch Service. Website. (2017). https://aws.amazon.com/elasticsearch-service/
[4] Amazon. 2017. Amazon EMR. Website. (2017). https://aws.amazon.com/emr/
[5] Amazon. 2017. Amazon Kinesis. Website. (2017). https://aws.amazon.com/kinesis/
[6] Amazon. 2017. Amazon QuickSight. Website. (2017). https://quicksight.aws/
[7] Amazon. 2017. Amazon Redshift. Website. (2017). https://aws.amazon.com/redshift/
[8] Amazon. 2017. AWS Data Pipeline. (2017). https://aws.amazon.com/datapipeline/
[9] Amazon. 2017. AWS Glue. Website. (2017). https://aws.amazon.com/glue/
[10] Amazon. 2017. Netflix and Amazon Kinesis Streams Case Study. Website. (2017). https://aws.amazon.com/solutions/case-studies/netflix-kinesis-streams/
[11] Amazon. 2017. Yelp Data Analytics Case Study. Website. (2017). https://aws.amazon.com/solutions/case-studies/yelp-data-analytics/
[12] Amazon. 2017. Zillow Provides Near-Real-Time Home-Value Estimates Using Amazon Kinesis. Website. (2017). https://aws.amazon.com/solutions/case-studies/zillow-zestimate/
[13] Jeffrey P. Bezos. 2015. Annual Letter to Shareholders. Press Release. (April 2015).
[14] Niraj Dawar. 2016. Use Big Data to Create Value for Customers, Not Just Target Them. Website. (Aug. 2016). https://hbr.org/2016/08/use-big-data-to-create-value-for-customers-not-just-target-them
[15] The Economist. 2017. Data is giving rise to a new economy. Website. (May 2017). https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy
[16] Synergy Research Group. 2016. AWS Remains Dominant Despite Microsoft and Google Growth Surges. Website. (Feb. 2016).
[17] Neil Hunt. 2016. Website. (2016). https://aws.amazon.com/solutions/case-studies/netflix/ Conference Presentation at AWS re:Invent 2016.
[18] Eric Knorr. 2016. Hot property: How Zillow became the real estate data hub. Website. (April 2016). https://www.infoworld.com/article/3060773/big-data/hot-property-how-zillow-became-the-real-estate-data-hub.html
[19] Ingrid Lunden. 2013. Amazon Takes Redshift, Its Cloud-Based Data Warehouse Killer, Global. Website. (Feb. 2013). https://techcrunch.com/2013/02/15/amazon-takes-redshift-its-cloud-based-data-warehouse-killer-global/
[20] Netflix. 2017. Investor Relations - Financial Statements. Website. (Sept. 2017). https://ir.netflix.com/
[21] David M. Search. 2010. mrjob: Distributed Computing for Everybody. Website. (Oct. 2010). https://engineeringblog.yelp.com/2010/10/mrjob-distributed-computing-for-everybody.html
[22] Softpedia. 2017. Netflix Users Spend 1 Billion Hours per Week Watching Movies. Website. (April 2017). http://news.softpedia.com/news/netflix-users-spend-1-billion-hours-per-week-watching-movies-514989.shtml
[23] Jeremy Stoppelman. 2013. Fast Company Innovation Uncensored. Panel Discussion. (Nov. 2013). http://blog.fastcompany.com/post/66283564254/yelp-ceo-jeremy-stoppelman-talks-big-data-in-this
[24] Laura Winig. 2016. GE's Big Bet on Data and Analytics. Website. (Feb. 2016). https://sloanreview.mit.edu/case-study/ge-big-bet-on-data-and-analytics/ Case Study.
[25] Steven Wu, Allen Wang, Monal Daxini, Manas Alekar, Zhenzhong Xu, Jigish Patel, Nagarjun Guraja, Jonathan Bond, Matt Zimmer, and Peter Bakas. 2016. Evolution of the Netflix Data Pipeline. Website. (Feb. 2016). https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905
[26] Yelp. 2017. Fact Sheet. Website. (June 2017). https://www.yelp.com/factsheet

Docker in Support of Big Data Applications and Analytics

Anand Sriramulu
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
asriram@iu.edu

ABSTRACT

To Discuss on the docker benefits in areas of *Continuous Deployment* and Testing, Security, Isolation, Multi-Cloud Platform and Environment Standardization, and explaining different use cases in which docker can improve the performance of Big Data applications.

KEYWORDS

i523, hid338, Data Science, Docker, Containers, Big Data Analytics, Cloud Computing

1 INTRODUCTION

Big data growing rapidly as the industries deal with large datasets in terms of Terabytes or Petabytes, in which we need for better solutions in the software development. Docker is a open source software containerization platform provides solutions for developers and sysadmins to build ship and run distributed applications whether on laptops, data centers,virtual machine or on the cloud. Docker provides lightweight environment that makes it easy to quickly deploy a piece of Software and the resources such as CPU, memory, disk, etc and make it portable and self-contained. [5]

2 DOCKER BENEFITS

Docker is a widely used container and it's being very matured compared to the others. I have outlined the top five benefits of using the ever-growing platform. [3]

2.1 Continuous Deployment and Testing

Continuous Deployment is a DevOps process in which the applications or features from continuous integration and deployed to production environment. Docker helps both development and devops and make sure it's consistency across environments.

Docker containers can be configured with all the configurations and dependencies internally, so that the developer need not worry about the environment as he can develop or test any product upgrades, maintenance releases, new features in a docker container and release the images to different production servers. This is a great advantage as it saves lot of time with no errors due to the deployment issues.[7]

2.2 Multi-Cloud Platforms

Docker being widely used container solution, all the cloud computing provides supports it, so the portability being greatest strength with docker. That means, the docker image running on AWS can be switched to Azure server easily as the applications built on the docker container is not depending on any platform. This gives the advantage for the application to free of Platform As a Service vendor lock and provides the level of abstraction from the infrastructure layer. List of hosting provides supporting docker includes AWS, Microsoft Azure, Digital Ocean, Exoscale, Google Compute Engine, OpenStack, Rackspace, IBM Softlayer, etc. [4]

2.3 Environment Standardization and Version Control

Docker support version control as like GIT or TFS repositories. The docker images can be version controlled and if there any issues in the deployment, it can be roll-backed to the previous version. The process of rollback is quick and easy when compared to VM backup and image creation processes.[7]

2.4 Isolation

Since docker is a container, it will be isolated from other containers and resources in the same host. Docker also make sure each container has it's own resource been allocated and isolated from the other containers. This gives the benefit on each container can run on its own application stack and be managed. So if an application is not needed, it can be removed by deleting the container and it will not leave any temporary or container related files on the host system. As mentioned earlier, each container has assigned with allocated resource, the docker make sure that it will not be exceeded. This prevents the issues related to the performance or down time of the other applications in the same host.[7]

2.5 Security

As the containers are isolated, Docker make sure that the applications that are running on containers have control only within their container. So no container can look into the processes of other container. Each container will have its own resources ranging from processing to network stacks which is great benefit as if there any impact to an application related to security it won't impact the other applications.

[8]

3 BIGDATA AND DOCKER

Big Data is one of the big trends in IT of recent years. Majority of companies are investing more time in collecting and managing data for the business needs. It's huge struggle for them to find a right system to get the relevant information needed for the business managers to make important decisions. Without big data, there are challenges to arm the organization with the technology stack, skilled professional and resources for the business intelligence to manage the data deluge. [10]

3.1 Use Docker To Avoid Dependency Issues

Each developer might have different set of big data tools and not to mention all the dependencies required, which then must be distributed to each machine in a cluster.

Companies assume this situation is manageable, but get enough developers on the same cluster and there are high possible chances for one tools requirements to break another. This will cause all the dependencies issues.

In this scenario, the companies either need to get the entire entire development team to use the common technology stack, or use Docker. Docker allows the developers to build each applications to be self contained with their dependencies. This gives the benefit to have different applications can run without a conflict.[11]

3.2 Reduce Reliance On MapReduce Experts With Pachyderm

Hadoop with MapReduce been popular for the distributed storage and big data framework. Pachyderm claims to be the modern Hadoop which uses Docker containers and Kubernetes for managing the clusters. Pachyderm Filesystem and Pachyderm Pipeplines are equivalent to HDFS and MapReduce respectivevly.

Hadoop relies on Java technology stack, in which it requires specialist programmers to build MapReduce job, in which Pachyderm gives the freedom for the programmers to choose any library and wrap it in the container for the data processing and integrate into the Pachyderm stack. [9]

3.3 Run Scheduled Analytics Using Containers With Chronos

The containers are a great way of deploying services at scale and giving isolation to services that run on the same host and improving utilization, but Docker can also be used for batch processing as well.

Chronos job scheduler provides a graphical user interface which allows the devops to run the Docker images into a Mesos cluster. This gives the benefit that the developers can use containers to run the scheduled data analytics.[12]

Chronos also gives the advantage that it doesn't need any manual setup on the cluster nodes to distribute the job processing in the containers.

3.4 Provision A Big Data Dev Environment Using Ferry

Ferry allows you to create big data clusters on the local machine (and AWS). The beauty of Ferry is that it allows anyone to define a big data stack using YAML, and then share it with other developers using a Dockerfile. As per the article [1], "Setting up a Hadoop cluster is as simple as:

backend:
- storage
personality: 'hadoop'
instances: 2
layers:
- 'hive'
Connectors:
- personality: 'hadoop-client'

Get started by typing

ferry start hadoop"

This will create a two node Hadoop cluster and a single Linux client. This can be customized at runtime or defined using a Dockerfile. Ferry is great for developers who want to get up and running with a big data environment using a test AWS box, developers that need a local big data dev environment, or users that want to share Big Data applications.

Running Ferry on AWS also has several advantages over something like Elastic MapReduce, such as not tying you to a single cluster of a single type (such as Hadoop).[1]

3.5 Run Big Data As Microservices With Coho

Large enterprise applications use publish-subscribe technology as part of the SOA Architecture can take the advantage of implementing each component as Microservice which solves a single purpose.

Microservice can help big data systems in terms of scalability(as each component runs independently), data quality(as data flows through the focused task to run the data analytics), Muliple Technology Stack.

Coho Data's DataStream Storage system provides the mechanism in which the docker containers can run on the data environment without any other needed infrastructure. The HDFS users can run the HDFS containers directly on the storage envrionment to extract the information needed for the data analytics[6]

4 CONCLUSIONS

The complex nature of big data and the tools used to analyze these data sets makes efficient processing difficult with standard environments.

While performance and pipeline efficiency were key components of this implementation, Docker containers also allow for application isolation from the host operating system. Since many big data tools have complex sets of dependencies and are difficult to build from source, the ability to deploy containers with different operating systems and dependency versions to the same host decreases the amount of effort needed to being analysis. With the use of containers allowed the deployment of each utility on its natively supported operating system, which improves stability and decreases the potential for dependency conflicts among software applications. There are other tools like Kubernetes or Docker Swarm can be used for container orchestration and helps us to distribute the containers in the clusters, but these applications will not provide application level workflows as it works within the container level. Additional implementation experience about the use of these tools within high-performance clusters may provide valuable insights about the scalability of these tools within data analytics workflows. Finally, the container technology is very helpful to deploy the applications to nearly any host system. While many factors can impact reproducibility, the use of containers limits variability due to differences in software environment or application configuration when appropriately deployed. The continued use of emerging technology and novel approaches to software architecture has the potential to increase the efficiency of computational analysis in big data. [2]

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.]. Big Data Development Environment using Docker. ([n. d.]). http://drydock.readthedocs.io/en/latest/
[2] [n. d.]. Containers and Orchestration Explained. ([n. d.]). https://www.mongodb.com/containers-and-orchestration-explained
[3] [n. d.]. Docker Benefits. ([n. d.]). https://opensource.com/resources/what-docker
[4] [n. d.]. Maintaining Docker Portability in a Multi-Cloud World. ([n. d.]). https://boxboat.com/2016/10/21/maintaining-docker-portability-multi-cloud-world/
[5] [n. d.]. What is Docker. ([n. d.]). https://www.docker.com/what-docker
[6] cohodata. [n. d.]. Coho. ([n. d.]). http://www.cohodata.com/pdfs/coho-data-solution-brief-microservices.pdf
[7] Sanket Dangi. [n. d.]. 5 Key Benefits of Docker: CI, Version Control, Portability, Isolation and Security. ([n. d.]). http://www.iamondemand.com/blog/5-key-benefits-of-docker-ci-version-control-portability-isolation-and-security/
[8] Sanket Dangi. [n. d.]. Docker in your Data: Data Services as Microservices. ([n. d.]). http://www.iamondemand.com/blog/5-key-benefits-of-docker-ci-version-control-portability-isolation-and-security/
[9] Susan Hall. [n. d.]. Pachyderm Open Source. ([n. d.]). https://thenewstack.io/pachyderm-aims-displace-hadoop-container-based-collaborative-data-analysis-platform/
[10] Tom Phelan. [n. d.]. Hadoop and Spark on Docker. ([n. d.]). https://www.bluedata.com/blog/2017/08/hadoop-spark-docker-ten-things-to-know/
[11] Manisha Sahasrabudhe. [n. d.]. Docker solved the depedency issues problem. ([n. d.]). https://dzone.com/articles/are-you-stuck-in-the-new-devops-matrix-from-hell
[12] Ken Sipe. [n. d.]. Containerize your batch jobs with Mesosphere and Docker. ([n. d.]). https://mesosphere.com/blog/docker-on-mesos-with-chronos/

What Separates Big Data from Lots of Data

Gabriel Jones
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
gabejone@indiana.edu

ABSTRACT

We briefly analyze the history of data to show how having *Lots of Data* hardly differs from data storage and analysis in the early days of SQL, or even before computers. We then explain how *Big Data* represents a paradigmatic shift from conventional data analysis. We then begin to look at the potential limits of *Big Data* to assert that this paradigmatic shift does not mean the end of science. We conclude that misunderstanding *Big Data* prevents organizations from capitalizing on its potential and can lead them to spurious answers.

KEYWORDS

i523, hid104, Big Data, Lots of Data, Data Science, Data History, Sociotechnical

1 INTRODUCTION

In 2008, Wired.com published an article titled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." They tried to assert that *Big Data*, at that point still a relatively new term, was such a revolutionary change that the scientific method would no longer exist [1]. Since at least 2008, professionals, scientists, and the public have flocked to the idea of *Big Data*, but many still struggle to understand both its grand potential and its realistic limits [5]. On one extreme, boasting terabytes of storage, they claim to be *Big Data* experts but only utilize *Lots of Data*, high quantities of traditional data. On the other extreme, they hyperbolize about how *Big Data* will change the world because it eliminates the need for educated hypothesizing, based on the fallacious assumption that *Big Data* is synonymous with *All Data*, implied by Wired.com and by some academics [3].

To avoid the common data deluge delusions, we borrow from a 2014 article written by Professor Carl Logoze, "Big Data, Data Integrity, and the Fracturing of the Control Zone," that defines *Big Data* as "data that disrupt fundamental notions of integrity and force new ways of thinking and doing to reestablish it" [3]. This definition both breaks the boundaries of *Lots of Data* and reins in the assumed panacea that leads people to believe they have *All Data*. Taking a brief look at the history of data clarifies what it means to have *Lots of Data*. A case study of the 1880s US Census Bureau demonstrates that mostly just volume and efficiency mark the difference between today's use of *Lots of Data* and the historical use of data in general, and how this differs from the definition and possibilities of *Big Data*. Having separated *Big Data* from an incorrectly limiting category, we then make the case for investigating what are the limits of *Big Data*. We briefly examine the basis for the argument that *Big Data* is not *All Data*, but a more rigorous analysis is beyond our current scope. We break down the first extreme, the synonymizing of *Big Data* with *Lots of Data*, by succinctly explaining how it represents a paradigmatic shift. We also hope to foster additional sociotechnical scholarly discussion and case studies of its limits, which would help break down the hyperbolic synonymizing of *Big Data* with *All Data*.

2 A BRIEF LOOK AT THE HISTORY OF DATA

The human ability to store and analyze data has evolved gradually over millennia. Although digital computer technology greatly accelerated this evolution, most mainstream uses of data still show signs of their historical roots. The formation of early libraries over 4,000 years ago signifies an important moment in methods of amassing data to be organized and processed by humans into knowledge [4]. Libraries still have prominence today both in the traditional sense, brick and mortar sites where one can study texts, and in a broader context, digital archives of algorithmically curated information. In either case, libraries are literal representations of information. If one wants to access a text, they can obtain a copy of it, physical or digital, and read the actual words of the text [3]. In contrast, another ancient technology, the abacus, demonstrates one of the first symbolic representations of data. The abacus uses an arrangement of beads to represent other numbers and calculations. The numbers themselves did not exist but were symbolically represented. This is an important early prerequisite to the emergence of statistics, which seeks to make accurate claims about a population based only on a sample [4].

One of the first uses of the term business intelligence, a feature of statistical analysis, was used in the *1865 Encyclopaedia of Commercial and Business Anecdotes*. The book described how a banker, Henry Furnese, gained an advantage over competitors by applying a structured method to collect and analyze information relevant to his business activities. Furnese's data analysis is considered one of the first of its kind for commercial purposes. It builds from the fundamental idea that the real world can be represented and analyzed symbolically, as a sample, to produce insights [4]. This is the same idea that allows, for instance, modern companies to provide performance bonuses to employees based on how well they meet certain criteria called Key Performance Indicators. It would be impossibly inefficient to have supervisors accurately observe every activity of every employee and objectively judge who made the most contributions, so instead, companies define metrics of good employee behavior and use these metrics to symbolically represent who adds the most value.

While being able to store and analyze data increased in importance near the end of the 19th century, the physical limits of storage and analysis, paper documents and human eyes, created a problem of *Lots of Data*. The US Census Bureau found themselves faced with this problem. As the US population skyrocketed, they estimated

that with late 19th century methods, it would take an estimated 8 years to process the data collected in the 1880 census. Processing the 1890s census data, they predicted, would take over 10 years, so it would not be ready to study until becoming outdated by the 1900 census. The solution came from a young engineer named Herman Hollerith, eventual founder of IBM and creator of the Hollerith Tabulating Machine. His machine mechanically processed punch cards so efficiently it that reduced 10 years of work to three months [4]. Thus, he effectively solved the problem of volume, processing data for the entire US population, and of efficiency, since a few machines successfully completed what would have taken countless human hours.

Overcoming the Census Bureau challenge marks a key moment in the history of dealing with *Lots of Data*. With the advent of digital computing and languages like SQL, technologies have continually risen to the ever-greater demands for volume and efficiency [2]. But armed with new technologies like web and mobile, society has created new types of relatively easily accessible data [4]. The inherently messy, unstructured, rapidly changing nature of this new data goes beyond what an abacus, a library, a Hollerith Machine, or a simple SQL database can handle. In addition to data volume and efficiency, *Big Data* introduces challenges of velocity, the unstable, constantly changing nature, and variety, the unification of datasets as distinct as website-eye mapping and social media network analysis [6]. This distinguishes itself from *Lots of Data*, a term whose significance depends mostly on perspective. Processing the census data used to be a challenge of *Lots of Data*, but with modern computing technology, storing and analyzing simple demographic data is relatively straightforward. *Big Data* offers no such historical asymmetry. Even as technology improves its capability of dealing with volume, the other factors that comprise *Big Data* will still pose challenges. In other words, a *Big Data* problem of yesterday is still a *Big Data* problem of today.

3 THE BEGINNING OF A NEW ERA, BUT NOT THE END OF SCIENCE

As the history of data shows, *Big Data* is not just a buzz word. It has real meaning that separates it from past notions of data; it represents a paradigmatic shift in the way we approach the representation and analysis of information, so much so that notions of integrity have been revisited. But this realization can easily be taken too far. In their book, titled *Big Data*, Mayer-Schonberger and Cukier go as far as providing an omniscient mathematical formula for *Big Data*, *(n = all)*, where *n* is the sample size and all is the population. They claim that *Big Data* represents all the data possibly available, with no limits on time, size, or variety, and therefore represents objective, absolute truth. The correlations we derive from *Big Data* therefore do not need proof of causation; the existence of a relationship or pattern in Big Data must be true of reality because *Big Data* is *All Data* [3].

While *Big Data* certainly does change the norms of what it means to prove causation, the *(n = all)* proposition falls short in theory and in practice. Numerous scholars argue that data, no matter what its size and complexity, is a sample, "with bias implicit due to choice of instrumentation, span of observation, units of measurement, and numerous other factors. In essence, *n* never equals all; all is a limit in mathematical terms that can be approached but never attained" [3]. Ignoring the implicit uncertainty of dealing with a data sample can provide misleading conclusions. The Google Flu Trends (GFT) provide an excellent example of over reliance on informal data and algorithmic models. GFT initially raised widespread scientific optimism; however, the predictions turned out to be highly exaggerated. Scholars that have analyzed the failure have acknowledge, among other factors, "an overconfidence in the veracity of the data as a true sample of reality, rather than a random snapshot in time and the result of algorithmic dynamics" [3]. The grand miscalculations of GFT should not have come as a surprise. Researchers have long-since understood the fallibility of data samples. *Big Data*, while opening up new possibilities for discovery of new questions, still must be held to standards of methodological credibility. Despite the hyperbolic optimism of the 2008 Wired.com article, scientific methods, theories, and ways of thinking will still play an important role in discovery.

4 CONCLUSIONS

As the latest development in the long history of data, *Big Data* represents a paradigmatic shift. *Big Data* clearly distinguishes itself from its predecessors in definition and in possibility. But, despite its tremendous, paradigm-shifting potential, *Big Data* is still an evolution on the long history of symbolic representation. Like any such representation, it shows but a small sample of the real world, viewed through the distorted lens of various biases. Adding the aspects of velocity and variety expand our avenues of discovery, but they do not eliminate the need for establishing some sense of scientific integrity, even if the norms of integrity must adapt. To be fair, the argument against the *(n = all)* proposition comes mostly from the scientific community, whose entire existence relies on integrity. The business world offers a different context. Often, decision-makers must take actions while relying on nothing more than structured but easily fallible methods of analysis. They try their best to produce reasonable insights with methods such as SWOT Analysis or Porter's Five Forces, because delivering timely, logical arguments often matters more than taking the time to find answers validated through scientific levels of scrutiny. In business, quickly finding reasonable answers often takes priority to slowly finding proven ones. Given their different priorities, businesses can perhaps afford to relax their standards of information integrity with *Big Data*, as long as they are cognizant of its inherent uncertainty. But it is this lack of cognizance that can lead people into the dangerous territory of making ill-advised decisions based on misleading data. In addition to clarifying what it means to go beyond *Lots of Data* to help people capitalize on the vast potential of *Big Data*, we hope to foster more sociotechnical research into what the dangerous territory of being incognizant looks like and how it can be avoided.

5 ACKNOWLEDGEMENTS

The author would like to thank Dr. Gregor von Laszewski and his teaching assistants for providing helpful feedback.

REFERENCES

[1] Chris Anderson. 2007. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Website. (June 2007). https://www.wired.com/2008/06/pb-theory/

[2] G. Groner and M. Rockwell. 1977. *Computer-Based Information Systems for a Hospital Emergency Department.* Technical Report. Rand Corporation, Santa Monica, Ca.
[3] 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
[4] Bernard Marr. 2015. A Brief History of Big Data Everyone Should Read. Website. (Feb. 2015). https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr/
[5] Bernard Marr. 2015. The Difference Between Big Data and a Lot of Data. Website. (Sept. 2015). http://data-informed.com/the-difference-between-big-data-and-a-lot-of-data/
[6] Svetlana Sicular. 2013. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. Website. (March 2013). https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#2826c61d42f6

Big Data and Analytics in Blockchain

Ashok Kuppuraj
Indiana University
Bloomington, Indiana 43017-6221
akuppura@iu.edu

ABSTRACT

Big data and its technologies help in augmenting and improving the current Blockchain technology and overcome the problems around it.

KEYWORDS

i523, hid324, Big data, Blockchain, Cryptocurrency, Bitcoin, Transaction, Smart chain

1 INTRODUCTION

The objective is to concur the abilities of the two broad topics in the current technology world, Big Data, and Block Chain. Blockchain and Big data are still evolving technologies, which gives us enough opportunity to explore and invent new concepts for its own good. As these are still evolving, we can leverage onefis solution on the other. To leverage each onefis problems and solutions, we must first identify the similarities in two frameworks and how these similarities are related and what solution we are going to adopt.

2 WHAT IS BIG DATA

Big data can be described as any type of data with large volume, velocity, and variety [4]. The history of Big data starts from the moment we started using computers back in the 1990s, however, we choose not to use all the generated data due to constraints in the processing and storage systems. Later, people understood that they are missing a lot of useful information to the business due to these constraints, and started leveraging data warehouse to process data in batch after data generation. At a certain point in time, even the data warehouse systems are not capable to handle the volume and velocity of data, we are generating[13] is exponential growth in data generation due to wide adoption of computers by humans in the form of mobile, PCs and introduction of IoT sensors, resulting in the need for technology to process these data and it is termed as ”Big Data” [14].

3 BLOCK CHAIN

Blockchain can be defined as a decentralized, public ledger persisted in a connected set of immutable Blocks. The core idea is to perform any set of a transaction without a governing third-party avoiding double spending by Distributed consensus. A transaction happens with an entity called tokens, tokens are the actual digital asset of a blockchain. The implementation begins with an entity A initiating the transaction, an initiated transaction request from A to B is broadcasted with Gossip protocol to most of the nodes, the transaction is validated by miners with the ledger available with them, the validation includes checking digital signatures and the previous input to that entity (i.e current withholding). Later the validated transactions are grouped with reference to its previous address and added as a current block. This block is then broadcasted to the network and the network peers validate the block and add them to their ledger, confirming the transaction. Hence, termed as ”Blockchain” [7].

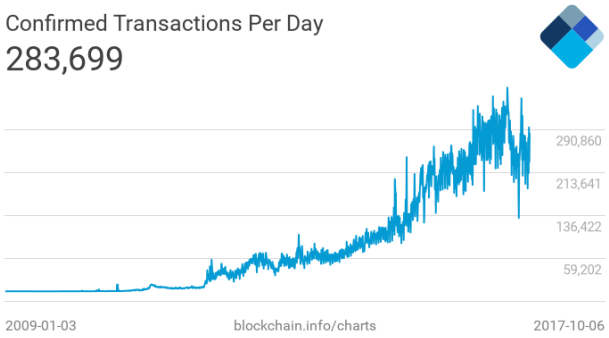

Figure 1: The snippet describes the number of transactions, in Bitcoin's network

4 BIG DATA VS BLOCKCHAIN

As far as data is concerned, both Big data and Blockchain go in parallel.Both involve processing data at volume, velocity, and variety which is the basic evaluation factor for defining big data. In the below section, the analysis is made on how these three V's corresponds to Block Chain, with an example from Bitcoin, one of the front-runners in implementing Blockchain technologies.

4.1 Data Volume

Though the data volume share of blockchain is considerably low compared to current Big data average, the volume it generates in an overall network perspective in terms of network I/O, logs, transaction data, it fits well with the terms of big data. For example, consider the transaction growth of Bitcoin [11], the volume of the transaction was averaging 5K in 2011, whereas in 2017 the average is 200K with an increase of 400 percent over 5 years and the volume is likely to grow in an exponential scale with the global acceptance of Blockchain technologies.

4.2 Data velocity

In data terms, even 10 MB of data is considered huge when its getting generated within a span of seconds, hence we must consider the velocity as an important metric in analyzing the data, here in Blockchain, though the transactions are not of high volume, but other non-token transactions like Gossip calls, smart contract transfers, block transfers, acceptance protocol were happening

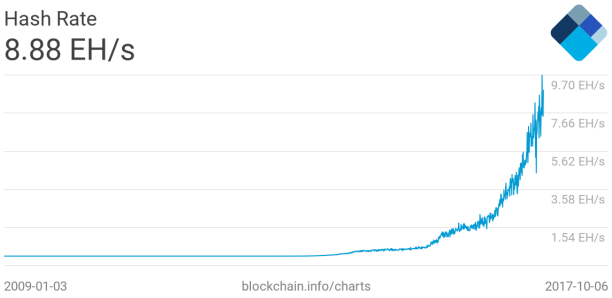

Figure 2: The snippet describes the number of hashes resolved per second, in Bitcoin's network [12]

every second, which in turns generate huge amount of data within 10 minutes of this interval with respect to Bitcoin.

4.3 Data Variety

In a wide perspective, Blockchain deals with multiple varieties of structured data like Token data, Smart contracts, consensus data, logging data and unstructured data like videos [10], audio depending upon the use-case of the Blockchain. And the popularity it has now and based on the current growth trend in the acceptance of decentralization, Blockchain technology will tend to generate more data in a wide range of varieties.

5 IMPLEMENTATION - BIG DATA TECHNOLOGIES IN BLOCKCHAIN

Before we start on the implementation of Big data technologies, its required to identify the problems Blockchain faces now in scope with Big data solutions, one of the main problems are slow transactions and visualization through complex analytic calculation. Though we have other problems, we consider the above two as the most important to enhance the success rate of this technology.For example, consider Bitcoin, the overall transaction timing is fewer in-terms of interbank transaction however for the end user it takes a minimum of 20 Min's to complete a transaction, whereas, in Visa, for instance, it can perform up to 24000 transactions per second [3].

5.1 Transaction processing

To deal with improving the transaction speed of a peer-to-peer network, first it is required to streamline the asynchronous process of gossip protocols, handshake between peers to increase the transaction processing speed, also by removing the block size limits[2] we can increase the frequency of the block building, this optimization can be easily achieved with the help of Big data queuing utilities like Apache Kafka by creating individual topics for each set of Broadcasting, Block Acknowledgement, and consensus sharing between peers, and second is to increase the Hash processing capacity by horizontal scaling with the help of Apache Spark or Apache Flink. By using these open source tools for hash processing, it is not required to invest on high-value GPU's to process data.

For an instance, Kafka can handle up to 200,000 messages/second (220MB/second)[8], which is way more than any other existing banking infrastructure can provide.

5.2 Visualization

The current visualization options available with blockchain is based on the shared ledger available in the network, to fetch the real-time reporting or visualizing the happenings in the network, one must have to take part in the network and share all the interactions and ledger details for any sort of analytic needs. As discussed in the previous section, if we start using Kafka for other peer-to-peer interactions via topics, we can seamlessly provide real-time reporting to users.

5.3 Smart-Blockchain

The next big leap in the Blockchain would be the implementation of Machine learning in the Blockchain. The current versions of blockchain don't have any machine learning modules or algorithm built along with. By including the machine learning modules in the blockchain network, Blockchain can be made smart by predicting malicious activities, optimizing transactions and evaluation of its sources.

5.4 Data Persistence

Storage is an important aspect of any platform, both in Big data and Blockchain, most of the data is persisted. In Blockchain, the data involving contracts and blocks are either stored in a file system or database [1], e.g Google's LevelDB in Bitcoin Blockchain. In Big data, the data storage is in Hadoop's File System or a database. Both use the data storage for write once and read many as their retention strategy. When it comes to data persistence, fault tolerance and recovery cannot be left behind. In big data technologies like Hadoop, the fault tolerance is ensured by HDFS, with the help of replication and Journal Managers. Whereas in Blockchain, the same has been ensured with Merkle tree data structure simulating Journal manager through validation and peer-to-peer network which simulates nodes of replication.

5.5 Decentralization

Decentralization can be defined as a distribution of functions or power[9], decentralization can be modeled in every stage of an application's lifecycle. In terms of data processing, data decentralization considers the data stays where it gets generated and the analytic happens at the same place. In large organizations, each unit independently generates data, process and analyze it without impacting the others. consider if it is a centralized system, the flexibility of each unit has to be constrained and output has to be generalized or standardized at the organization level, this seriously impacts the evolving needs of each units[5]. Hence decentralization can be a good approach in order to cope with the changing world. However, the changes cannot be easily incorporated with conventional technologies like ETL (Extract-Transform-Load) tools and mainframe which runs based on fixed terms, the big data technologies come into rescue with the schema-less data model, distributed file system, etc.

Both Big data technologies and Blockchain technologies go hand in hand with decentralization. Let's consider in a view of data processing, In Apache Hadoop framework the data is processed locally in the individual nodes whereas in Informatica the data is transferred to centralized servers to perform any processing[6]. In

Blockchain, the hash processing happens in the peer nodes instead of a centralized server and shared with the other peers for validation and acceptance.

6 CONCLUSION

Although Blockchain provides a solution for Real life problems, it would be nearly impossible without its implementations leaning towards Big data solutions. Big data and its technologies is a front-runner in the handling of data of different volume, velocity, and variety which Blockchain is yet to reach. With the current acceptance rate of Blockchain, Big data, and Machine learning technologies, maybe in future, countries don't need a leader to take decisions on their behalf, people can collectively take a state decision and election process will be so simple that it can happen every day.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Andreas M. Antonopoulos. 2014. *Mastering Bitcoin: Unlocking Digital Crypto-Currencies* (1st ed.). O'Reilly Media, Inc., 1005 Gravenstein Highway North Sebastopol, CA 95472 USA.

[2] James Hudon. 2017. Dear Bitcoin, This is How You Can Beat Visa fi?! James Hudon fi?! Medium. https://medium.com/@hudon/dear-bitcoin-this-is-how-you-can-beat-visa-b5ee857cf193. (Aug 2017).

[3] VISA Inc. 2010. Transaction Analysis. IBM. (8 2010).

[4] Doug Laney. 2001. *3D Data Management : Controlling Data Volume, Velocity, and Variety.* Technical Report. META Group.

[5] Julie Lockner. 2015. Is Data Decentralization the New Trend? — Sandhill. http://sandhill.com/article/is-data-decentralization-the-new-trend/. (Aug 2015).

[6] David Meister. 2012. *Informatica - 9.x - Informatica Architecture: Nodes and Domains - (English).* Technical Report. Informatica Corporation.

[7] Satoshi Nakamoto. 2008. *Bitcoin: A Peer-to-Peer Electronic Cash System.* Technical Report. bitcoin.org. 9 pages. https://bitcoin.org/bitcoin.pdf

[8] Brock Noland. 2015. Apache Kafka Performance Numbers fi?! phData. https://www.phdata.io/apache-kafka-performance-num. (2015).

[9] Editors of the American Heritage Dictionaries. 1988. *American Heritage Dictionary of the English Language* (5 ed.). Houghton Mifflin Harcourt Publishing Company, 125 High Street, Suite 900 Boston, MA 02110.

[10] Doug Petkanics. 2017. Introducing Livepeer fi!? A Decentralized Live Video Broadcast Platform and Crypto Token Protocol. https://medium.com/@petkanics/introducing-livepeer-a-decentralized-live-video-broadcast-platform-and-crypto-token-protocol-7eb4b1de47ed. (Mar 2017).

[11] BLOCKCHAIN LUXEMBOURG S.A. 2017. Confirmed Transactions Per Day - Blockchain. https://blockchain.info/charts/n-transactions?timespan=all. (2017).

[12] BLOCKCHAIN LUXEMBOURG S.A. 2017. Hash Rate - Blockchain. https://blockchain.info/charts/hash-rate?timespan=all&showDataPoints=true. (2017).

[13] SAS Institute Inc. 2016. What Is Big Data? — SAS US. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html. (2016).

[14] Asaf Yigal. 2017. The Exponential Growth of Data. https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/. (Feb 16 2017).

Creating Better Urban Environments with Optimized Public Bus Routes and Schedules

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

ABSTRACT

Optimized public bus networks reduce greenhouse gases while providing a safe and affordable way to travel. Fighting against an unglamorous reputation and stereotypes of inconvenience, modern bus networks must continually update their routes and schedules to meet the demands of modern commuters. Fortunately, big data analytical methods can identify optimal routes through human mobility mining and optimize schedules through dynamic coordinating bus clusters and station and inter-station controls. These methods make it more possible than ever to offer a dynamic and convenient public bus network. Fully optimized public bus systems have the potential to increase quality of life standards, catalyze new development, and prevent urban sprawl.

KEYWORDS

i523, hid225, bus route optimization, bus schedule optimization, public bus transit

1 INTRODUCTION

Public transit systems are at the core of every urban environment. Between 2005 and 2015 public transportation has grown 15% compared with a 5% growth for private vehicle travel [36]. Often seen as a second-class transportation method, public bus networks are at the foundation of American public transportation systems. In most urban areas, bus transit is the only form of public transportation. In the United States and Canada there are 83 public rail networks including, heavy-rail, light-trail, subways, and streetcars. In contrast, there are 1,018 public bus networks [36]. Accordingly, public bus transport accounted for 49.1% of all passenger trips and 37.6% of passenger miles [36].

Optimizing the reliability and usability of these transit systems is essential. The United Nations estimates 87% of Americans will live in urban environment by 2050 compared to 82% in 2017 [8]. In a competitive market to attract the next generation of high-paying jobs and talented workforces, successful urban areas will leverage their optimized transportation systems.

1.1 Benefits of Public Transportation

1.1.1 Agglomerative Economics. Agglomerative economics, also known as urban increasing returns, is the theory that higher urban density increases labor market pooling, input sharing, and knowledge spillovers [39]. Research shows that efficient and abundant public transportation increases agglomerative economics [24]. Thus, investing in public transportation is critical to successful urban planning. Studies show that an increase in public transit investment yields higher per capita GDP and average wages. In large urban areas, a 10% increase in transit investment can create up to $1.8 billion in agglomeration economic benefits [11]. For example, the city of Cleveland recently invested $50 million in bus rapid transit (BRT) which spurred $5.8 billion in new transit oriented development [21].

Other benefits created from increased density from better transportation options include shorter commutes and people with shorter commutes report systematically higher subjective well-being [42]. Health benefits also exist with higher urban density because higher density promotes more active transportation methods and physical activity [37] [20] [18].

1.1.2 Cost Savings. In the top 20 metro areas by public transportation ridership, individuals using public transportation as their primary transit method saved $10,064 annually [12]. Public transit riders save money on car payments, maintenance, insurance, fuel, and parking expenses.

1.1.3 Environmentally Friendly. "The Nobel Prize winning 2007 Intergovernmental Panel on Climate Change report concluded that greenhouse gas emissions must be reduced by 50% to 85% by 2050 in order to limit global warming to four degrees Fahrenheit [23]." Optimizing public bus systems can play a major role in reducing greenhouse gas emissions. Compared to private vehicle transportation, average bus transit occupancy reduces carbon dioxide emissions per passenger mile by 33.33%, but when bus transit is fully occupied, carbon dioxide emissions per passenger mile are reduced by 81.25% [23]. Public buses will further reduce fuel and greenhouse gases as new electric and hybrid-electric buses are introduced [29]. In addition, compared with personal vehicle travel heavy commuter rail and light rail reduced greenhouse emissions per mile by 76% and 62% respectively [1].

Public transportation reduces energy consumption further by limiting the need for vehicle transportation infrastructure, manufacturing new vehicles, and extracting more fossil fuels [1].

1.1.4 Safety. According to the National Safety Council, 40,200 people died in traffic accidents in 2016 [7]. In the same time period, 4.6 million people were seriously injured from traffic accidents in America [25]. A report by the American Public Transit Association in association with the Victoria Transport Policy Institute found that public transportation is 10 times safer per passenger mile than private vehicle transit [31].

1.1.5 Citizen Health. Public transit riders have better mental and physical health than their car driving peers. Researchers at the University of East Anglia found that active methods of transportation improved commuters mental well being [34]. 76% of Americans commute by car alone [38]. On the other hand, public transportation allows riders to destress, relax, read, and socialize

[34]. Physical health is improved through active transportation methods like walking, cycling, and even public transportation. Public transit commuters get three times the amount of daily exercise than those that drive [27]. In fact, public transportation commuters spend roughly 25 minutes a day walking to and from their stops [33]. As a result, Body Mass Index scores were reduced for new public transit riders [9] [30]. In addition, public transit also increases air quality and reduces pollution.

1.1.6 Reduced Traffic Congestion. Public transit relives traffic congestion on the most congested roads. During the 35 day 2003 Los Angles transit strike, traffic increased 47% on average and nearly 100% along the most popular transit routes [4]. Residents served by public transportation saved 865 million hours in commute time [10]. In fact, the largest form of mass transit in America the school bus industry removes nearly 36 cars from the road each day [15].

2 OPTIMIZATION TECHNIQUES

Bus ridership numbers are falling across America, in Los Angles bus ridership dropped 8.9%, in New York City bus ridership dropped 16% between 2002-2015, and in 2015 Chicago had 25 million less bus boardings than 2013 [14]. According to research group TransitCenter, customer satisfaction relies on service frequency and travel times, not modern and flashy amenities like free wifi and power outlets. In fact, transit riders desire improved station conditions, real-time information, and service reliability instead of 21st century upgrades many transit systems are funding [44]. In one recent example, the Metropolitan Transportation Authority in New York City announced plans to add 2,042 high-tech buses with both wifi and power outlets costing nearly $5,000 extra per bus, totalling over $10 million [32]. Instead of investing in these superfluous upgrades, we suggest using this money to optimize current routes and schedules.

2.1 Schedule Optimization

Bus bunching is a common phenomenon where buses on the same route arrive at the same station at the same time. Bus bunching occurs because the loading time of the first bus is longer than the second bus. At each stop, the second bus loads passengers faster until the two buses converge at the same stop [5].

Headway distribution, the time between bus arrivals, is a major measure of service quality and reliability. Interestingly, passengers would rather headway regularity rather than scheduled punctuality [26]. Preventing bus bunching while maintaining short wait times is the primary objective of bus scheduling. Delgado [17] organized bus bunching and scheduling techniques into three operational categories: station control, inter-station control, and capital rearrangement.

2.1.1 Station Control. Station control techniques include static and schedule based holding [5], dynamic holding [6], stop skipping [43], and boarding limits [46]. Mazloumi [35] used ant colony and genetic algorithms to optimize bus transit schedules.

2.1.2 Inter-Station Control. Inter-station control techniques include controlled bus cruising speeds [22], bus overtaking [40], and transit priority signal mechanisms [3].

2.1.3 Capital Rearrangement. Alternatively, bus systems can add buses at the beginning of the route or even in the middle, but this is inefficient us of drivers and buses for both the bus systems and their passengers [5]. If bus bunching on a specific route is predictable, one of the above optimization techniques should be used to limit the effect.

In 2012, the San Francisco Municipal Transportation Agency implemented the first all door boarding policy, allowing passengers to board from both the front and back door [2]. Traditionally, buses only allow front door boarding to prevent fare avoidance, but a two year review of San Fransico's all door boarding policy actually shows an increase in fare compliance in combination with faster trips and short board times [1].

2.2 Route Optimization

Historically, public transportation systems use human surveys to understand people's transportation needs. Despite the substantial time and cost spent on the survey process, the macroscopic analysis based on surveys is too static to reflect the fast development of urban areas [28]. As a result, many transportation networks still use routes from out-dated studies and surveys.

2.2.1 Route Consolidation. Route consolidation in Portland increased bus route speed by 6% without sacrificing passenger perceived quality [19]. The TransitCenter explains the how American bus stops are too close together. In New York City, the average distance between bus stops is only 750 feet. In fact, buses in New York City spend 22% of their active time at bus stops [45]. They suggest consolidating bus stops to combine unnecessary stops that slow the everyone's ride. Creating new stops within a quarter mile (a 5 minute walk or less) prevents isolating existing riders [45].

2.2.2 Human Mobility Mining. Significant research shows the predictability of human mobility patterns. Montjoye [16] showed that four spatio-temporal points are enough to identify 95% of individuals [28]. Song [41] shows that human mobility has a predictability of 93%. Made easier but the constant data collection from private vehicle transportation like taxis and rider sharing apps, transportation systems can use this data to create better routes to meet their passengers needs. Liu [28] used data from 30 million taxi trips to optimize bus routes in Beijing. Chuah [13] used taxi data in Singapore to identify public transportation islands and proposed new routes to serve passengers in these areas.

3 CONCLUSION

Amid a major disruption in the transportation industry due to technological advances in autonomous vehicles, public bus networks must adapt. For instance, public bus networks can leverage this new technology and create a larger network of smaller autonomous buses that are highly optimized to local human mobility patterns. Using live traffic data, the Internet of Things, current events, and fast computing big data algorithms, this advanced system of public transportation could eliminate the need for private vehicle ownership all together.

Public transportation and bus transit in particular are vital components of urban environments. Maximizing their effectiveness

through optimization of routes and schedules will insure their importance in the urban landscape.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Federal Transit Administration. 2016. Transitfis Role in Environmental Sustainability. (May 2016). https://www.transit.dot.gov/regulations-and-guidance/environmental-programs/transit-environmental-sustainability/transit-role

[2] SFMTA Municipal Transport Agency. 2014. *All-Door Boarding Evaluation Final Report.* resreport. SFMTA Municipal Transport Agency.

[3] Eric Albright and Miguel Figliozzi. 2012. Analysis of the Impacts of Transit Signal Priority on Bus Bunching and Performance. In *Proceedings of the Conference on Advanced Systems for Public Transport.* CASPT, Santiago, Chile, 1–11.

[4] Michael L. Anderson. 2014. Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion. *American Economic Review* 104, 9 (September 2014), 2763–2796. https://ideas.repec.org/a/aea/aecrev/v104y2014i9p2763-96.html

[5] Matthias Andres and Rahul Nair. 2017. A predictive-control framework to address bus bunching. *Transportation Research Part B: Methodological* 104, Supplement C (2017), 123 – 148. https://doi.org/10.1016/j.trb.2017.06.013

[6] John J. Bartholdi and Donald D. Eisenstein. 2012. A self-coordinating bus route to resist bus bunching. *Transportation Research Part B: Methodological* 46, 4 (2012), 481 – 491. https://doi.org/10.1016/j.trb.2011.11.001

[7] Neal E. Boudette. 2017. U.S. Traffic Deaths Rise for a Second Straight Year. (Feb. 2017).

[8] Bret Boyd. 2017. Urbanization And The Mass Movement Of People To Cities. Internet. (01 2017). https://graylinegroup.com/urbanization-catalyst-overview/

[9] Barbara B. Brown, Carol M. Werner, Calvin P. Tribby, Harvey J. Miller, and Ken R. Smith. 2015. Transit Use, Physical Activity, and Body Mass Index Changes: Objective Measures Associated With Complete Street Light-Rail Construction. *American Journal of Public Health* 105, 7 (2015), 1468–1474. https://doi.org/10.2105/AJPH.2015.302561 PMID: 25973829.

[10] Nicholas Brown. 2013. Public Transportation Saved 865 Million Hours Of Delay On US Roads In 2011. (2013). https://cleantechnica.com/2013/02/08/public-transportation-saved-865-million-hours-of-delay-on-us-roads/

[11] Daniel G. Chatman and Robert B. Noland. 2014. Transit Service, Physical Agglomeration and Productivity in US Metropolitan Areas. *Urban Studies* 51, 5 (2014), 917–937. https://doi.org/10.1177/0042098013494426

[12] Chad Chitwood. 2014. August Transit Savings Report Shows Individuals Save $10,064 a year. Press Release. (Aug. 2014). http://www.apta.com/mediacenter/pressreleases/2014/Pages/140814_Transit-Savings.aspx

[13] Seong Ping Chuah, Huayu Wu, Yu Lu, Liang Yu, and Stephane Bressan. 2016. Bus Routes Design and Optimization via Taxi Data Analytics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16).* ACM, New York, NY, USA, 2417–2420. https://doi.org/10.1145/2983323.2983378

[14] Josh Cohen. 2017. As Bus Ridership Declines, Transit Experts Make The Case For All-Door Boarding. (Feb. 2017). https://nextcity.org/daily/entry/bus-ridership-declines-experts-make-case-for-all-door-boarding

[15] American School Bus Council. 2017. Environmental Benefits. (2017). http://www.americanschoolbuscouncil.org/issues/environmental-benefits

[16] Yves-Alexandre de Montjoye, Csar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3 (March 2013), 12. https://doi.org/10.1038/srep01376 Article number: 1376.

[17] Felipe Delgado, Juan Mufloz, Ricardo Giesen, and Aldo Cipriano. 2009. Real-Time Control of Buses in a Transit Corridor Based on Vehicle Holding and Boarding Limits. *Transportation Research Record: Journal of the Transportation Research Board* 2090 (2009), 59–67. https://doi.org/10.3141/2090-07

[18] Delfien Van Dyck, Greet Cardon, Benedicte Deforche, James F. Sallis, Neville Owen, and Ilse De Bourdeaudhuij. 2010. Neighborhood SES and walkability are related to physical activity behavior in Belgian adults. *Preventive Medicine* 50, Supplement (2010), S74 – S79. https://doi.org/10.1016/j.ypmed.2009.07.027

[19] Ahmed El-Geneidy, James G. Strathman, Thomas J. Kimpel, and David Crout. 2005. Effects of Bus Stop Consolidation on Passenger Activity and Transit Operations. *Transportation Research Record* 1971 (11 2005), 1–22.

[20] Ulf Eriksson, Daniel Arvidsson, Klaus Gebel, Henrik Ohlsson, and Kristina Sundquist. 2012. Walkability parameters, active transportation and objective physical activity: moderating and mediating effects of motor vehicle ownership in a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity* 9, 1 (05 Oct 2012), 123. https://doi.org/10.1186/1479-5868-9-123

[21] Instituite for Transportation and Development Policy. 2013. More Development for Your Transit Dollar: An Analysis of 21 North American Transit Corridors. (Nov. 2013). https://www.itdp.org/more-development-for-your-transit-dollar-an-analysis-of-21-north-american-transit-corridors/

[22] Sheng-Xue He. 2015. An anti-bunching strategy to improve bus schedule and headway reliability by making use of the available accurate information. *Computers & Industrial Engineering* 85, Supplement C (2015), 17 – 32. https://doi.org/10.1016/j.cie.2015.03.004

[23] Tina Hodges. 2010. *Public Transportationfis Role in Responding to Climate Change.* Technical Report. U.S. Department of Transportation Federal Transit Administration. https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/PublicTransportationsRoleInRespondingToClimateChange2010.pdf Editing/Design: Jarrett Stoltzfus.

[24] Joseph Jenkins, Michael Colella, and Frederick Salvucci. 2011. Agglomeration Benefits and Transportation Projects. *Transportation Research Record: Journal of the Transportation Research Board* 2221 (2011), 104–111. https://doi.org/10.3141/2221-12

[25] Kirsten Korosec. 2017. 2016 Was the Deadliest Year on American Roads in Nearly a Decade. (Feb. 2017). http://fortune.com/2017/02/15/traffic-deadliest-year/

[26] J. Lin and M. Ruan. 2009. Probability-based bus headway regularity measure. *IET Intelligent Transport Systems* 3, 4 (December 2009), 400–408. https://doi.org/10.1049/iet-its.2008.0088

[27] Todd Litman. 2010. *Evaluating Public Transportation Health Benefits.* resreport. Victoria Transport Policy Institute. http://www.apta.com/resources/reportsandpublications/Documents/APTA_Health_Benefits_Litman.pdf

[28] Yanchi Liu, Chuanren Liu, Nicholas Jing Yuan, Lian Duan, Yanjie Fu, Hui Xiong, Songhua Xu, and Junjie Wu. 2017. Intelligent Bus Routing with Heterogeneous Human Mobility Patterns. *Knowl. Inf. Syst.* 50, 2 (Feb. 2017), 383–415. https://doi.org/10.1007/s10115-016-0948-6

[29] Marcy Lowe, Bengu Aytekin, and Gary Gereffi. 2009. *Public Transit Buses: A Green Choice Gets Greener.* Environmental Defense Fund, 257 Park Avenue South New York, NY 10010. http://proxyiub.uits.iu.edu/login?url=https://search-proquest-com.proxyiub.uits.iu.edu/docview/58827750?accountid=11620 Date revised - 2010-02-03; Publication note - Environmental Defense Fund, 2009; Last updated - 2016-09-28.

[30] John M. MacDonald, Robert J. Stokes, Deborah A. Cohen, Aaron Kofner, and Greg K. Ridgeway. 2010. The Effect of Light Rail Transit on Body Mass Index and Physical Activity. *American Journal of Preventive Medicine* 39, 2 (Aug. 2010), 105–112. https://doi.org/10.1016/j.amepre.2010.03.016

[31] Paul Mackie. 2016. Transit is 10-times safer than driving fi?! and makes communities safer, says new APTA report. (Sept. 2016). https://mobilitylab.org/2016/09/08/transit-10-times-safer-driving-makes-communities-safer-says-new-apta-report/

[32] Thomas MacMillan. 2016. MTA to Roll Out New Buses With Wi-Fi, Phone-Charging Outlets. (March 2016).

[33] Jason Margolis. 2015. Why taking the bus is better for our health than driving. (Oct. 2015). https://www.pri.org/stories/2015-10-28/why-taking-bus-better-our-health-driving

[34] Adam Martin, Yevgeniy Goryakin, and Marc Suhrcke. 2014. Does active commuting improve psychological wellbeing? Longitudinal evidence from eighteen waves of the British Household Panel Survey. *Preventive Medicine* 69, Supplement C (2014), 296 – 303. http://www.sciencedirect.com/science/article/pii/S0091743514003144

[35] Ehsan Mazloumi, Mahmoud Mesbah, Avi Ceder, Sara Moridpour, and Graham Currie. 2012. Efficient Transit Schedule Design of timing points: A comparison of Ant Colony and Genetic Algorithms. *Transportation Research Part B: Methodological* 46, 1 (2012), 217 – 234. https://doi.org/10.1016/j.trb.2011.09.010

[36] John Neff and Matthew Dickens. 2016. *2016 PUBLIC TRANSPORTATION FACT BOOK.* resreport 67th Edition. American Public Transportation Association, 1300 I Street, NW, Suite 1200 East Washington, DC 20005. http://www.apta.com/resources/statistics/Documents/FactBook/2016-APTA-Fact-Book.pdf

[37] Neville Owen, Ester Cerin, Eva Leslie, Lorinne duToit, Neil Coffee, Lawrence D. Frank, Adrian E. Bauman, Graeme Hugo, Brian E. Saelens, and James F. Sallis. 2007. Neighborhood Walkability and the Walking Behavior of Australian Adults. *American Journal of Preventive Medicine* 33, 5 (2007), 387 – 395. https://doi.org/10.1016/j.amepre.2007.07.025

[38] Clara Reschovsky. 2003. *Journey to Work: 2000.* resreport. U.S. Census Bureau. https://www.census.gov/prod/2004pubs/c2kbr-33.pdf

[39] Stuart S. Rosenthal and William C. Strange. 2004. Chapter 49 - Evidence on the Nature and Sources of Agglomeration Economies. *Handbook of Regional and Urban Economics* 4 (2004), 2119–2171. https://doi.org/10.1016/S1574-0080(04)80006-3

[40] Jan-Dirk Schmocker, Wenzhe Sun, Achille Fonzone, and Ronghui Liu. 2016. Bus bunching along a corridor served by two lines. *Transportation Research Part B: Methodological* 93, Part A (2016), 300 – 317. https://doi.org/10.1016/j.trb.2016.07.005

[41] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021. https://doi.org/10.1126/science.1177170

arXiv:http://science.sciencemag.org/content/327/5968/1018.full.pdf
[42] Alois Stutzer and Bruno S. Frey. 2008. Stress that Doesn't Pay: The Commuting Paradox*. *Scandinavian Journal of Economics* 110, 2 (2008), 339–366. https://doi.org/10.1111/j.1467-9442.2008.00542.x
[43] Aichong Sun and Mark Hickman. 2005. The Realfi?!Time Stopfi?!Skipping Problem. *Journal of Intelligent Transportation Systems* 9, 2 (2005), 91–109. https://doi.org/10.1080/15472450590934642
[44] TransitCenter. 2016. *Whofis On Board 2016.* resreport 2. TransitCenter, TransitCenter One Whitehall Street 17th floor New York, NY 10004. https://transitcenter.org/publications/whos-on-board-2016/#introduction
[45] TransitCenter. 2017. Bus Stop Balancing. (Oct. 2017). http://transitcenter.org/2017/10/30/bus-stop-balancing/
[46] Shuzhi Zhao, Chunxiu Lu, Shidong Liang, and Huasheng Liu. 2016. A Self-Adjusting Method to Resist Bus Bunching Based on Boarding Limits. *Mathematical Problems in Engineering* 2016 (May 2016), 7. https://doi.org/10.1155/2016/8950209 Article ID 8950209.

Big Data Applications and Autonomous Vehicles

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

ABSTRACT

We will explain the importance of autonomous vehicles, Big Data applications used on these vehicles, and several computational methods used for achieving successful autonomy.

KEYWORDS

i523, HID343, Autonomous Vehicles, Safety, Neural Networks, Analytics in Autonomous Vehicles

1 INTRODUCTION

The way of life is changing every day with the help of technological advancements. We keep hearing more and more by companies how they are trying to make the computers to think and react like human beings. Autonomous vehicles are one of the products of these advancements.

The main difference of learning between humans and computer is the way of gathering experience. Humans learn from experience but computers learn from the data. This is why data is the most fundamental aspect for the computer to do given task. We show about how computers process data what kind of analytic used for processing and learning from data and importance of big data.

2 IMPORTANCE OF AUTONOMOUS VEHICLES

Autonomous vehicles are essential, and it is the future of driving method to going A to B. There are several reasons for it which will change the future of driving. Before we go into detail about autonomous vehicles we need to learn types of autonomous vehicles which are listed below:

Level 0(no automation): The driver responsible for all aspects of vehicle instruments while monitoring road conditions[7].

Level 1(at least one automation): This level, automation needs to have at least one function to be automated, and functions need to be independent if there is more than one automated function[7].

Level 2(combined-function automation): This level requires a minimum of two automated functions to perform a task for the driver[7].

Level 3(limited self-driving automation): The driver, must have control of all the safety-related features under certain conditions while the vehicle is monitoring changes[7].

Level 4(full self-driving automation): At level 4 vehicle will monitor conditions of the environment and perform all the critical driving actions for the driver[7].

2.1 Safety Aspect

Increasing safety features in motor vehicles is decreasing the number of crashes, if we look at the data from U.S. Census Bureau we can see that from 1980 to 2009 we have decreased in accidents. We can not tie all these decreases into technology because road structure and personal education also increased. According to the National Highway Traffic Safety Administration(NHTSA) and as shown in Figure 1 and Figure 2,

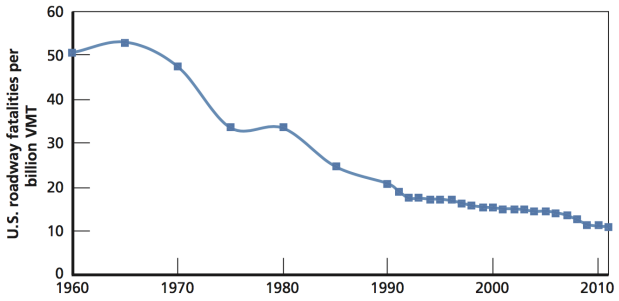

Figure 1: This data from BTS(2013) includes all highway transportation

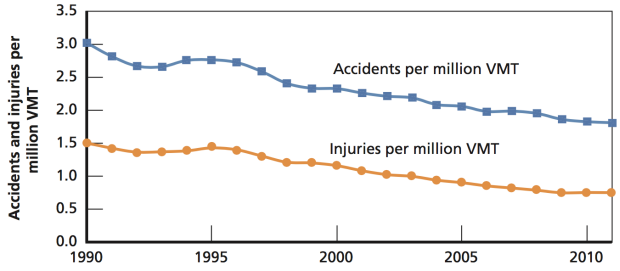

Figure 2: This accident report includes all highway transportation and every crash which involved two vehicles counted as one.

- The number of motor vehicle accidents that happened in 2010 valued at 242 billion which represents 1.6 % of Gross Domestic Product[3].
- There were 32,999 deaths and 13.6 million injured individuals from accidents that occurred in 2010[3].

2.2 Economic Aspect

Beside from saving money from increasing the safety while decreasing the economic cost of motor vehicles, autonomous vehicles can

improve many aspects of business and government supply chain industry, gas usage, time of commuting, and productivity.

According to RAND Corporation research, benefits of autonomous vehicles which includes productivity, gas consumption, increased safety aspects, improved mobility outperforms the disadvantages of autonomous vehicles[1].

2.2.1 Fuel Consumption. As technology improves every day, we see live traffic events in our navigation apps, and optimally its steering individuals to go to different directions based on traffic events to eliminate any waste from commuting[12]. With the help of autonomous vehicles, this live data directly go to intelligent vehicle systems and an autonomous car will steer their directions without the need of human interaction. An intelligent system like this improve the fuel consumption and decrease the commuting time from point A to B[12]. As shown in Figure 3 fuel consumption gain relative to improvements in autonomous vehicles[12].

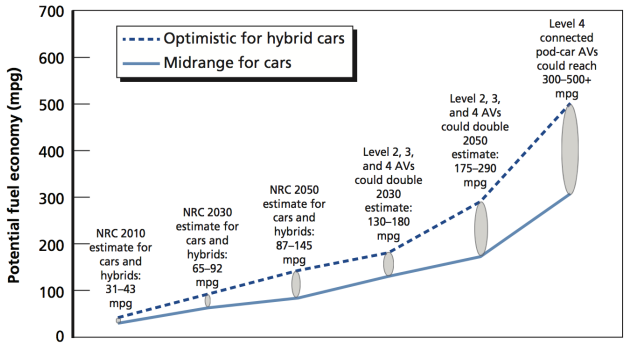

Figure 3: Potential fuel economy difference between hybrid, conventional, and autonomous cars.

2.2.2 Supply Chain. As of year, 2017 majority of companies essential success to stay in competitive is the supply chain [5]. Autonomous vehicles will have a high impact on supply chain distributions because of ability to operate 24/7 in right circumstances, and faster travel times. Self-driving vehicles will also help the current driver shortage situation in supply chain businesses [5]. As shown in forecast of truck driver shortage in Figure 4 [5].

There was approximately 38,000 truck driver shortage in 2014. This value expected to increase and reach 175,000 by 2024 [5].

2.2.3 Productivity. Autonomous vehicle will also give people to do multitasking abilities for productivity improvements. Individuals will have more free time to do other tasks.

It is reported that currently, every driver spends average one hour on travelling[10]. This time could use to more productive work with the expected self-driving technology. [10].

3 HOW IT RELATED TO BIG DATA AND WHAT DOES BIG MEAN?

Data can come from various resources. In our case, it is sensors, signals, cameras, customer behaviors and many others resources [2]. This data can be structured and unstructured dependent on

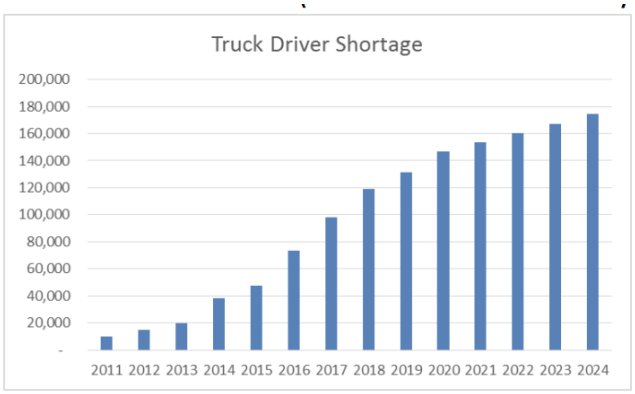

Figure 4: Forecast of truck driver shortage

where it is coming from while the term BIG refers to the volume of data it may also refer to techniques and tools that have been used to process this significant amount of data this tools can vary from cloud computing, visualization techniques to artificial intelligence procedures for analyzing[2].

The essential success of autonomous vehicles depends on data. The more data they have the correct decision can the autonomous vehicles do. As we stayed before this data comes from the variety of places some of them are sensors, GPS signals, cameras, internet connectivity [2]. All this data helps the car to make intelligent decisions while analyzing those data, without the data it will never successfully reach to the destination[9].

Additionally, companies are using "big data to optimize customer experience and operational safety ultimately laying the groundwork for the fully autonomous vehicles[2]." Companies can get collect data about customer driving habits by integrating additional sensors to its cars[8]. This connectivity to Big Data platform can give companies advantages over deploying new features to their cars. One another importance of Big Data connectivity for autonomous cars is the ability to transfer learning experience to other autonomous vehicles in other words when one autonomous vehicle learns from data and road conditions then that data can be transferable millions of other autonomous vehicles in contrast to individual experience which stays with the person [8]. Please see Figure 5 and Figure 6 about how connectivity implementation happens in Big Data [8].

4 ANALYTICS USED ON AUTONOMOUS VEHICLES

In this topic we will examine some of the methods that used in autonomous vehicles and these will include Machine Learning, Deep Learning, Artificial Intelligence.

4.1 Machine Learning

Machine learning widely used for many applications. Some of this applications include image and voice recognition, spam detection, fraud detection, the stock market, teaching a computer how to play chess, and, off course self-driving cars.

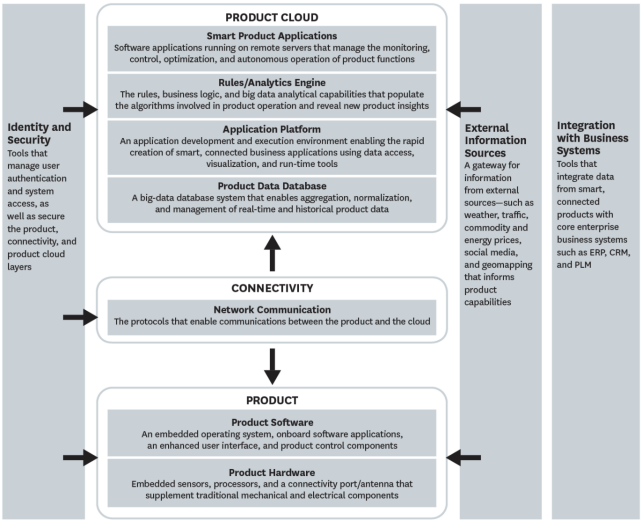

Figure 5: How big data and connectivity works

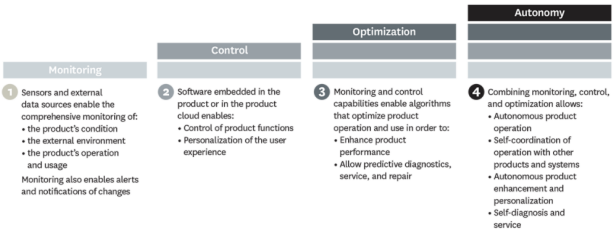

Figure 6: Process of connectivity

Machine learning is teaching computers to learn to perform a task from past experiences this experience comes from data. Self-driving cars equipped with ECU (Electronic Control Units). These ECUs process data from sensors like Lidar, radars, cameras or the IoT(Internet of Things) and they are equipped with machine learning algorithms to make decisions in different conditions[9]. These decisions vary from adjusting the speed with different driving conditions to recognizing the pedestrian movement on the road. Please see Figure 7 for understanding the world from an autonomous vehicle perspective[6].

4.1.1 Should we store this data in someplace or analyze it simultaneously. Current technology and Big Data methods allows self-driving or any other autonomous vehicles to analyze data on the go[11].

4.2 Conventional Neural Networks

Conventional Neural Networks (CNN)[4] used in pattern recognition applications. The significant advantage of CNN is that it can automatically learn features of the data from training examples[4]. This gives the significant advantage over learning features from image recognition. The image comes from camera system mounted on a car, after capturing images they will go through CNN, and

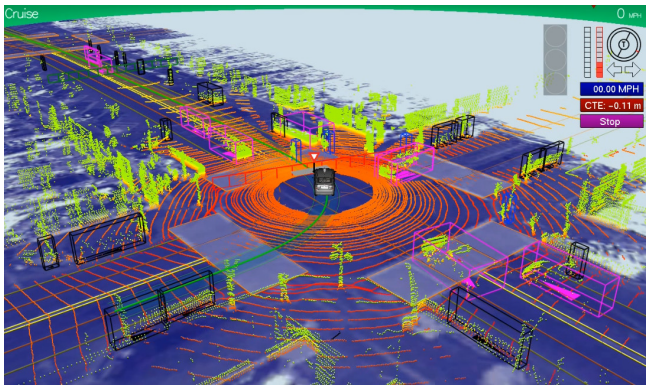

Figure 7: World from an eye of autonomous vehicle

after recognizing the features on the road, it will give the vehicle to steer itself based on computed steering command[4]. As shown in Figure 8 CNN model process steps[4].

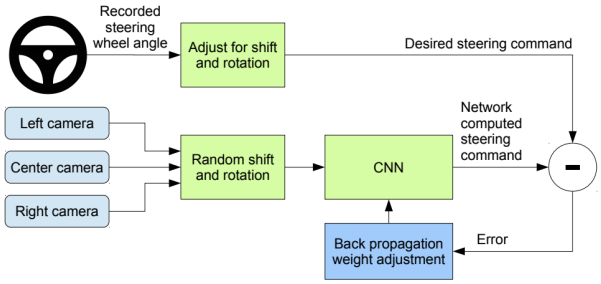

Figure 8: CNN process steps

4.3 Big Data for Predicting Safety Road Passage

Big data can help to maximize safety aspects in self-driving cars by using Big Data mining and analytics. This kind of analytics will require vehicle and analytics to connect in cloud-based systems also this will require an entirely automated car, in this case, it is Level 4 which is a fully integrated self-driving car [7]. This autonomy will give the vehicle to the ability to choose the safe passage at all times automatically as shown in Figure 10 [7].

The system still requires a driver to turn on the car and put the car id. After that, it requires the driver to put the destination. When Big Data engine receives all the required input, it will start predictions for road segments based on real-time Big Data analysis. If the cloud system does not predict any accidents in that road segments than vehicle continuous it is the destination as usual if the cloud system predicts any accidents then it reroutes the vehicle path to the destination[7].

This kind of cloud system will also give the user to ability choose between fastest route or the best fuel consumption, but the safety is always going to be the first priority. Figure 9 and Figure 10 shows the pseudo-code for implementing in Big Data Engine[7].

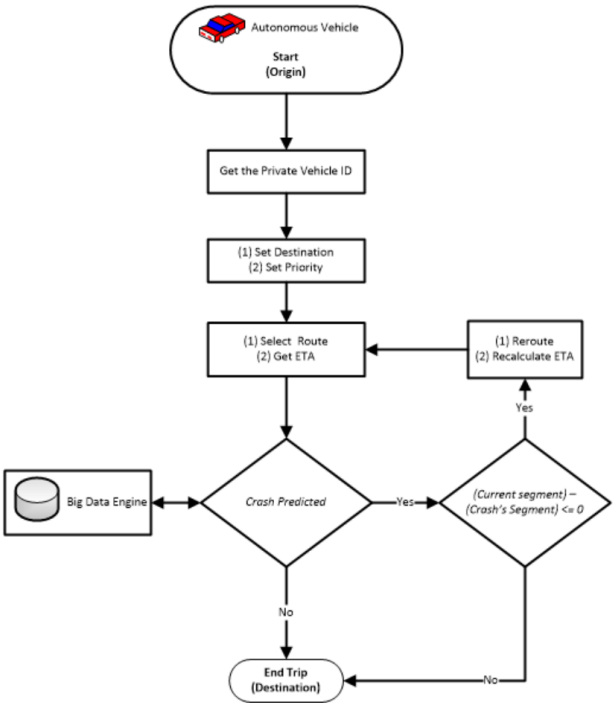

Figure 9: Big data road passage safety process steps

Algorithm 1: Check the Predicted Accident Location

initialization;
if *(CurrentSegment#) - (CrashSegment#)*

$$\leq 0$$

then
  (1) Reroute;
  (2) Recalculate ETA;
  (3) Recalculate Distance;
  (4) Recalculate Fuel Consumption;
else
  Keep using the current trajectory ;
end

Figure 10: Pseudo-code for road passage safety

5 CONCLUSION

We showed importance of autonomous vehicle and Big Data based applications on an autonomous vehicle is presented. Insights about advantages of autonomous vehicles had given. Several analytical approaches while using Big Data applications had shown.

ACKNOWLEDGMENTS

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

REFERENCES

[1] M. J. Anderson, Karlan N., K. Stanley, D., P. Sorensen, C. Samaras, and O. Oluwatola, A. 2016. *Autonous Vehicle Technology.* Technical Report. RAND Corporation. https://doi.org/10.7249/RR443-2
[2] Vengie Beal. n.d.. Big Data. Online. (n.d.). http://www.webopedia.com/TERM/B/big_data.html
[3] L. Blincoe, T. Miller, E. Zaloshnja, and B. Lawrence. 2010(Revised). *The Economic and Societal Impact of Motor Vehicle Crashes, 2010.* techreport DOT HS 812 013. National Highway Traffic Safety Administration.
[4] M. Bojarski, D. Testa, D., D. Dworakowski, B. Firner, B. Flepp, P. Goyal, D. Jackel, L., M. Monfort, U. Muller, J. Zhang, X. Zhang, and J. Zhao. 2016. *End to End Learning for Self-Driving Cars.* techreport. Nvidia Corporation, NVIDIA Corporation Holmdel, NJ 07735. https://arxiv.org/pdf/1604.07316v1.pdf
[5] B. Costello and R. Suarez. 2015. *Truck Driver Shortage Analysis.* Technical Report. American Trucking Associations.
[6] Li De. 2017. Google X: Leveraging data and algorithms for self-driving cars. (2017). https://digit.hbs.org/submission/google-x-leveraging-data-and-algorithms-for-self-driving-cars/
[7] Florida State University 2016. *Autonomous Vehicles Safe-Optimal Trajectory Selection Based on Big Data Analysis and Predefined User Preferences.* Florida State University. https://doi.org/10.1109/UEMCON.2016.7777922
[8] E. P. Michael and E. H. James. 2014. How Smart, Connected Products Are Transforming Competition. Web Page. (Nov. 2014). https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition
[9] Savaram Ravindra. n.d.. The Machine Learning Algorithm Used in Self-Driving Cars. (n.d.). http://www.kdnuggets.com/2017/06/machine-learning-algorithms-used-self-driving-cars.html
[10] M. Sivak and B. Schoettle. 2016. *Would Self-Driving Vehicles Increase Occupant Productivity?* techreport SWT-2016-11. University of Michigan, The University of MichiganSustainable Worldwide Transportation 2901 Baxter RoadAnn Arbor, Michigan 48109-2150 U.S.A.
[11] Statistical Analysis System. n.d.. The Connected Vehicle: Big Data, Big Opportunities. (n.d.).
[12] David Ticoll. 2015. *Driving Changes: Automated Vehicles in Toronto.* Technical Report. University of Toronto.