Use Cases in Big Data Software and Analytics

Vol. 1, Fall 2017

Bloomington, Indiana

Sunday 8th October, 2017, 01:07

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

Contents

Chapter 1

Preface

1.1 List of Papers

Name	HID	Title
hid101	Huiyi Chen	Big Data and standardize testing
hid102	Dianprakasa, Arif	This is my paper about xyz
hid104	Jones, Gabriel	What Separates Big Data from Lots of Data?
hid105	Lipe-Melton, Josh	This is my paper about data visualization in sports
hid106	Qiaoyi Liu	Big Data Analytics in Groceries Stores
hid107	Ni,Juan	This is my paper about xyz
hid109	Shiqi Shen	Big Data in Social Media
hid111	Lewis, Derek	Big Data Analytics in Biometric Identity Management
hid201	Arnav, Arnav	Big Data analytics and Edge Compting
hid202	Himani Bhatt	Big data analytics in Weather forecasting
hid203	Chandwani, Nisha	Big Data Analytics using Spark
hid204	Chaturvedi, Dhawal	Big Data Anaytics and High Performance Computing
hid205	Chaudhary, Mrunal L	Applications of Big Data in Fraud Detection in Insurance
hid208	Devineni, Jyothi Pranavi	This is my paper about Big Data and Deep Learning
hid209	Han, Wenxuan	Big Data Application in Web Search and Text Mining
hid210	Hotz, Nicholas	Natual Language Processing of Electronic Health Records
hid211	Ajinkya Khamkar	Distributed environment for neural network
hid212	Kumar, Saurabh	Big Data Analysis using MapReduce
hid213	Liu, Yuchen	Big Data and Speech Recognition
hid214	Lu, Junjie	Big Data and Basketball
hid215	Mallala, Bharat	Big Data and Artificial Neural Networks
hid216	Millard, Mathew	Big Data Analytics in Sports - Track and Field
hid218	Niu, Geng	Big data's influence on e-commerce and lifestyle
hid219	Parampali Sreenath, Syam Sundar Herle	Big Data Analytics Architecture for Real-Time Traffic Control
hid224	Rawat, Neha	Big Data Applications in the Hospitality Sector
hid225	Schwartzer, Matthew	Optimizing Mass Transit Bus Routes with Big Data
hid228	Swargam, Prashanth	Big data applications in Electric Power Distribution
hid229	ZhiCheng Zhu	Big Data and Machine Learning
hid230	YuanMing Huang	Big data with natural language processing
hid231	Vegi, Karthik	Using Big Data for Fact Checking
hid232	Rahul Velayutham	Big Data Analytics in Sports - Soccer
hid233	Wang, Jiaan	Big Data Applications in Media and Entertainment Industry

hid234	Weixuan Wang	Big Data Analytics in Tourism Industry
hid235	Wu, Yujie	Big Data in Recommendation System
hid236	Yang Weipeng	Big Data in MOOC
hid237	Ahmed, Tousif	Big Data Analytics in Cyber Security and Threat Research
hid301	Arora, Gagan	Big Data Analytics in Finance Industry
hid302	Sushant Athaley	Big Data Application in Restaurant Industry
hid303	Brunetti Nademlynsky, Lisa	This is my paper about xyz
hid304	Ricky Carmickle	Big Data and Astrophysics
hid305	Andres Castro Benavides	Big Data Analytics for Municipal Waste Management.
hid306	Cheruvu, Murali	The Internet of Things and Big Data
hid308	Pravin Deshmukh	Big Data and Data Visualization
hid309	Dubey, Lokesh	BigData Analytics using Apache Spark in Social Media
hid310	Kevin Duffy	Big Data Applications in Food Insecurity
hid311	Durbin, Matthew	Big Data and Healthcare
hid312	Neil Eliason	An Overview of Big Data Applications in Mental Health Treatment
hid313	Tiffany Fabianac	Big Data Platforms as a Service
hid314	Fadnavis, Sarang	Big Data analytics in Media industry
hid315	Garner, Jeffry	Roles and Impact on Mobility Network Traffic in Big Data
hid316	Robert Gasiewicz	Big Data Analytics in Biometric Identity Management
hid318	Irey, Ryan	This is my paper about xyz
hid319	Mani Kumar Kagita	Big Data Analytics for Municipal Waste Management
hid320	Elena Kirzhner	Big Data Analytics and Applications in Childbirth
hid321	Knapp, William	This is my paper about xyz
hid323	Uma M Kugan	This is my paper about NoSQL Databases in support of Big Data Applications and Analytics
hid324	Ashok Kuppuraj	Big data in Blockchain
hid325	J. Robert Langlois	Impact of Big Data on the Privacy of Mental Health Patients
hid326	Mahendrakar, Mohan	Bigdadta in Clinical Trails
hid327	Marks, Paul	Using Big Data to minimize Fraud, Waste, and Abuse (FWA) in United States Healthcare
hid328	Dhanya Mathew	Big data analysis in Finance Sector
hid329	Ashley Miller	Big Data Analytics in Higher Education Marketing
hid330	Janaki Mudvari Khatiwada	Big data in Improving Patient Care
hid331	Tyler Peterson	Big Data Applications In Population Health Management
hid332	Judy Phillips	Big Data Analytics in Agriculture
hid333	Anil Ravi	Big Data and Artificial Intelligence solutions for In Home, Community and Territory Security
hid334	Peter Russell	AWS in support of Big Data Applications and Analytics
hid335	Sean Shiverick	Big Data Analytics, Data Mining, and Public Health Informatics: Using Data Mining of Social Media to Track Epidemics
hid336	Jordan Simmons	Recommendation Systems on the Web
hid337	Ashok Reddy Singam	Big Data and Artificial Intelligence Solutions for in Home, Community and Territory Security
hid338	Sriramulu, Anand	Docker in support of Big Data Applications and Analytics
hid339	Hady Sylla	Big data application for treatment of breast cancer
hid340	Tim Thompson	Big data analytics for archives and research libraries
hid341	Tibenkana, Jacob	This is my paper about xyz
hid342	Udoyen, Nsikan	Big data analytics in college football (NCAA)
hid343	Usifo, Borga	Big Data Applications in Self-Driving Cars (Approval Waiting)

hid345	Wood, Ross	Big data analytics in the entertaiment industry.
hid346	Zachary Meier	Big Data in Oceanography
hid347	Jeramy Townsley	Sociological Applications of Big Data
hid348	Budhaditya Roy	Using Singularity for Big Data

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

i523, 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 [? ?]. 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 [?]. 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[?]. 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 [?]. 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 [? ? ?]. 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 [?]. 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 [?]. 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 [?]. 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 [?]. 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 [? ?]. 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 [?]. 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 [?]. 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 [?]. 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 [?].

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 [?]. 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 [?]. 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 [?]. 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.

REFERENCES

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

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

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

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

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

Big Data’s influence on ecommerce and lifestyle

Geng Niu
Indiana University Bloomington
752 Woodbridge Dr
Bloomington, Indiana 47408
gengniu@iu.edu

ABSTRACT

Big data has become the buzz words in recent years and it exerts huge influence on e-commerce and our lifestyle. However, for the general public big data is still something mysterious. This paper will serve as a review of what ways big data is utilized to improve e-commerce and influence our daily life.

KEYWORDS

big data, ecommerce

1 INTRODUCTION

People were used to get dressed well at weekends and drove or took public transportation to the centers of cities or towns to choose what they like in physical stores. However, this is never necessary with the rapid development of e-commerce driven by internet technologies and better logistics. E-commerce became a buzz word about 15 years ago but it came into being in 1991 when internet started to be used for commercial purposes. “At first, the term e-commerce meant the process of execution of transaction electronically. In 2000, the word e-commerce was redefined as the process of purchase of available goods and services over the internet.” [1] This paper will focus on big datafis influence on the newer definition of e-commerce instead of the first one which is very broad.

2 THE COMING OF THE BIG DATA ERA

In the traditional mode of commerce, consumers need to go to physical stores and take time to look for products they want by walking. Companies manufacturing these products do commercials on TV and newspapers to attract potential consumers. This mode of doing business did not change in the beginning of e-commerce, and the difference is that sellers moved into virtual shops from a real shop. In the web 2.0 era, search engines enabled consumers to look for products in virtual shops and sellers can receive feedback in their website. [4]

However, in the mobile and sensor-based era e-commerce is drastically changed. “The number of mobile phones and tablets (about 480 million units) surpassed the number of laptops and PCs (about 380 million units) for the first time in 2011” [4] The wide spread of mobile devices and other sensor-based devices enables the gathering of huge volume of data which is fresher and more accurate compared with data gathered from surveys and questionnaires. “In most cases, e-commerce firms deal with both structured and unstructured data. Whereas structured data focuses on demographic data including name, age, gender, date of birth, address, and preferences, unstructured data includes clicks, likes, links, tweets, voices, etc.”[2] With huge and various data available and relevant technologies, the big data era came.

3 MONITORING CONSUMERS’ JOURNEY IN ONLINE TRANSACTION

Big data analytics makes more data driven strategies for businesses to reach their consumers. With the use of data generated from Electronic Data Interchange, business runners can gain better understanding of consumer behavior so as to improve customer service and business strategies. Customers can be labeled into different segments or groups according to the patterns of their purchase online with their demographic information. By doing so, customers can be easily targeted especially during campaigns and festival sales because companies invest a lot to attract customers and retain existing base.[8] For example, Amazon is providing more customized offers, advertisements and discounts to consumers because it can identify patterns in consumers’ shopping habit which is enabled by analyzing cookies and clickstream on consumer browsers.[5]

There are two technologies associated with the monitoring of consumers’ journey online. One is text mining which relies on the use of text-based content from blogs and social media sites. Based on the information obtained, judgments on relevant issues can be made.[5] Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output.[11] Another technology is sentiment analysis which is based on learning algorithm or artificial intelligence to make clear about attitudes to a particular good or service. The words obtained from the data will be will analyzed and tagged and then are interpreted whether the opinion is positive or not.[5]

4 PERSONALIZED SERVICES

Recommender systems or recommendation systems are also used by e-retailers to provide personalized services. Recommender system is a subclass of information filtering system that seeks to predict the “ratin” or “preference” that a user would give to an item. Recommender systems can generate recommendations in two ways. The first way is called collaborative filtering, which means a model from a user’s past shopping behaviors including purchase records and ratings to given items as well as decisions made by other users will be built to generate recommendations. Another approach to build model is the content-based filtering. In this approach, a series of discrete characteristics of an item will be used to recommend additional items with similar properties.[7]

Let’s take Amazon as an example. 35% of Amazon’s revenue can be attributed to its recommendation engine. Amazon has on-site recommendations. When users click “your recommendations” link, they will see the products that the system recommends to them. Another way to recommend products is through the “frequently bought together”. For instance, when a user is searching for a laptop he or she wants to buy, he or she probably sees that a backpack which can hold the laptop is recommended. Some other ways of providing recommendations is “your browsing history”, “related to items you’ve viewed” and personal emails.[6]

5 DYNAMIC PRICING

When customers are shopping online, there is an electronic seller bargaining with you. This technology is called dynamic pricing. “Some business set different prices for their products or services based on algorithms that take into account competitor pricing, supply and demand and other external factors in the market. It is a common practice in industries such as hospitality, travel, entertainment, retail, electricity and public transport.”[10] Amazon customers can receive different or customized prices or discounts for the same item. This is set by the resource planning system through the use of data of previous purchase, clicksteam, cookies and so on. CNN once reported that for a particular DVD, the price increased by $ 2.5 after the customer deleted his cookies.[5] Here are some strategies for dynamic pricing. The first one is high-value customer price which means that customers who always pay full price for a certain product or service rarely get information about promotion or discounts. Another strategies is based on demand and supply and the time. For example, a seller could price up the products like a coat in extreme weather because it is possibly very needed by people.[3]

Although this function has its benefits such as increasing margin profits, customers may view it as price discrimination. However, there are some differences between dynamic price and price discrimination. Price discrimination happens when a sell changes the price of a product or service according to a consumer’s demographics. In contrast dynamic price more focuses on price fluctuations in demand and competitive landscape.[9]

6 LIFE CHANGES

Lifestyle changes are caused by a combination of internet, mobile devices and e-commerce. And big data is not possible to apply without the advances made in internet and ecommerce. For business runners, big data provides them more opportunities to gain profits. They are more likely to locate a potential buyer of their products and services. For example, the recommendation system can help them find customers not only in the local area but also in other cities or even countries. An seller who operate his store in Taobao in Shanghai receives a large number of orders from customers all over China in holidays, which is not imaginable. However, they are facing pressure from the increased workload. Since customers do not have to go to a physical store, they may order their goods whenever online. In Taobao, sellers are supposed to be online for most of the day. Even a customer ordered something at 10.pm, the seller should send notification of the order and answer questions the customer has. Another example is the way people watch dramas and movies. In the past, one had to wait in front of the TV and found no good dramas. Or he went to the cinema only to find the movie not worth the money for the ticket at all. On with Netflix and Youku, a Chinese video site, one can know millions of other viewersfi rating and choose the types of dramas and movies they like. It is certain that big data in e-commerce is influencing people’s lifestyle shopping, traveling, eating and entertainment.

7 CONCLUSION

With the spread of mobile phones and laptops and affordable high-speed internet, people now have their electronic assistants. Amazon and Taobao will tell you what products that you may be interested in are available. Google will send you notification about news that you have been following. When you are on the street, you mobile phone may tell you what are the good restaurants nearby. It has no doubt that big data has made people’s life more convenient. And big data renders business runners more opportunities to gain profits as well as more competition from other people.

REFERENCES

[1] 2008. History of Ecommerce. (2008). http://www.ecommerce-land.com/history_ecommerce.html

[2] Shahriar Akter and Samuel Fosso Wamba. 2016. Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets* 26 (2016).

[3] Matthew Bertulli. 2017. 5 Dynamic Pricing Strategies for eCommerce Growth. (2017). https://www.demacmedia.com/dynamic-pricing-strategies-ecommerce/

[4] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. (2012).

[5] Uyoyo Zino Edosio. 2014. Big Data Analytics and its application in E-commerce. (2014).

[6] Tom Krawiec. 2017. The Amazon Recommendations Secret to Selling More Online. (2017). http://rejoiner.com/resources/amazon-recommendations-secret-selling-online/

[7] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com Recommendations Item-to-Item Collaborative Filtering. (2003).

[8] Durjoy Patranabish. 2016. How Big Data is impacting the e-commerce industry. (2016). https://yourstory.com/2016/11/big-data-impacting-e-commerce-industry/

[9] Brian Smyth. 2015. Dynamic Pricing and Price Discrimination: Whatfis the Difference? (2015). http://blog.wiser.com/dynamic-pricing-and-price-discrimination-whats-the-difference/

[10] Wikipedia. 2017. Dynamic Pricing. (2017). https://en.wikipedia.org/wiki/Dynamic_pricing

[11] Wikipedia. 2017. Text mining. (2017). https://en.wikipedia.org/wiki/Text_mining

Big Data Applications in the Hospitality Sector

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

ABSTRACT

This paper focuses on how big data is used in the hotel industry for better customer satisfaction, marketing effectiveness and yield management using customer data for segmentation and predictive analyses.

1 CONCLUSIONS

This is the conclusion.[1]

ACKNOWLEDGMENTS

Acknowledgements

REFERENCES

[1] Gregor V Laszewski. 2017. test. (2017).

Big Data Analytics in Tourism Industry

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

ABSTRACT

This paper focuses on how the tourism industry has been impacted by the development of the Internet and improvements in information and communication technologies and how big data analytic can influence tourism research.

KEYWORDS

Big data analytics, tourism

1 INTRODUCTION

this is my introduction [1].

2 CONCLUSIONS

This my conclusion.

ACKNOWLEDGMENTS

The authors would like to thank I523.

REFERENCES

[1] 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)*. 5–7. https://doi.org/10.1109/BigData.2014.7004475

Big Data inRecommendation 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

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

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

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?[5]

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 paper.

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. Refer figure 1.

[Figure 1 about here.]

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[11]. 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.[6]

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 etc, along with family information is also an unstructured data. Weather and traffic information also constitutes as an important data to consider. [6]

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. [6]

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"[9]. 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. 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[9]. 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[6],[9]. Guest check analytics can help determine what sells well, how often somebody orders certain items and detailed pricing analyses[6]. 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[6]. 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[12]. 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[8]. 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[8].

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.[7][1]

MyCheck and MarketingVitals.com together provide mobility and data analytics platform for the hospitality industry. [3]

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. [10]

Upserve, a restaurant management platform, provides payment processing, point of sale, data insights to boost margins and exceed guest expectations.[12][2]

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.[6]

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.[6]

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.[6]

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.[10]

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[12]. 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[4]. 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.[7]

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.

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

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.

REFERENCES

[1] [n. d.]. ([n. d.]). https://www.fishbowl.com

[2] [n. d.]. ([n. d.]). https://upserve.com

[3] 2015. (Sept 2015). http://www.businesswire.com/news/home/20150916005807/en/MyCheck-Marketing-Vitals-Announce-Integration-Big-Data

[4] 2015. Big Data's Last Crusade: Restaurants still slow to embrace smart technology. *FastCasual.com* (May 2015).

[5] 2016. Restaurant industry to navigate continued challenges in 2016. (02 2016). http://www.restaurant.org/News-Research/News/Restaurant-industry-to-navigate-continued-challeng

[6] National Resturant Association. 2014. Big Data and Restaurants: Something to Chew On. Web. (11 2014). https://www.restaurant.org/Downloads/PDFs/BigData

[7] 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

[8] LISA JENNINGS. 2015. Making big data small. *Nation's Restaurant News* 49, 7 (May 2015), 22–23.

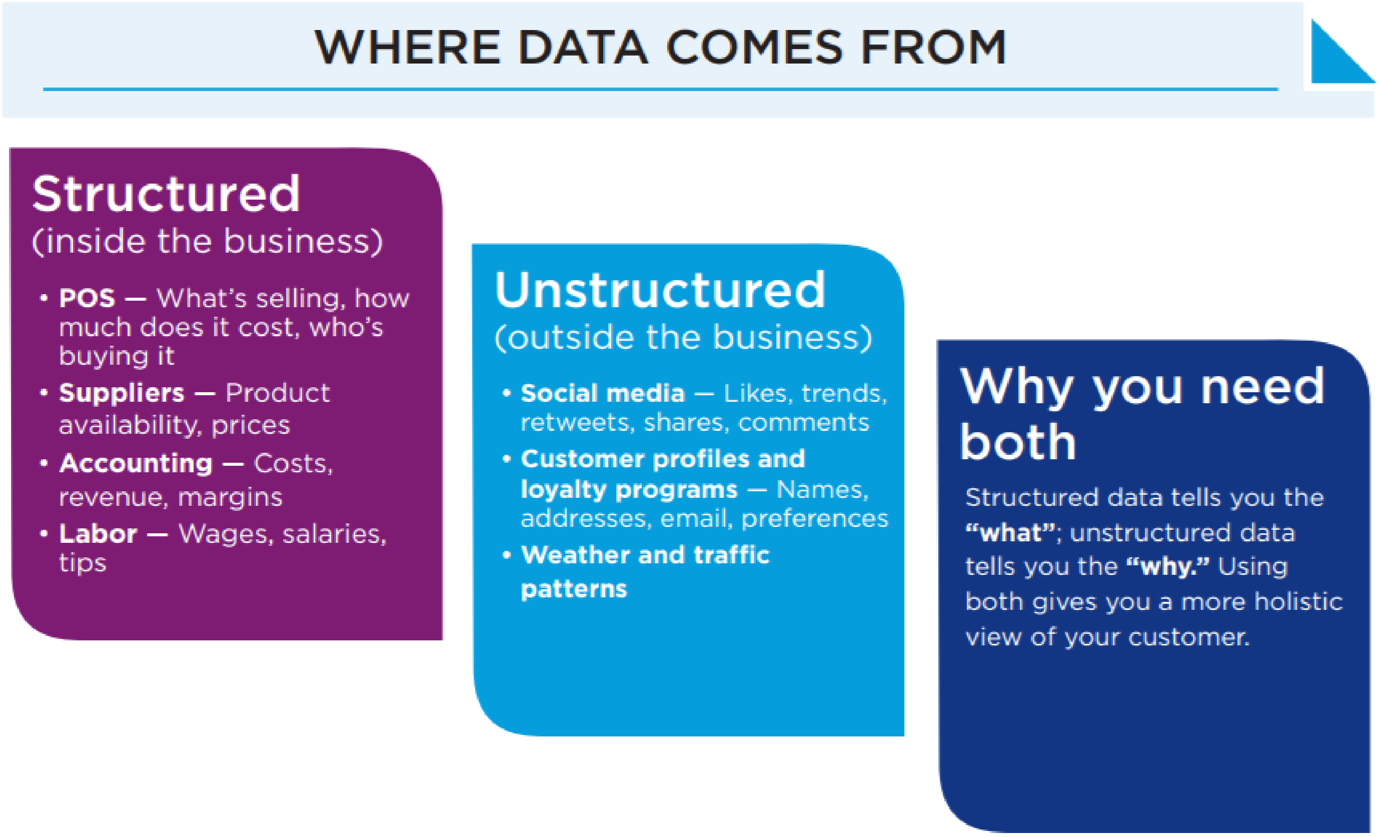
[9] Amanda C. Kooser. 2013. BIG DATA. *Restaurant Business* 112, 9 (September 2013), 24–31.

[10] 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.

[11] John Morell. 2013. Get a Grip on Big Data. (may 2013). https://www.qsrmagazine.com/operations/get-grip-big-data

[12] 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

List of Figures



WHERE DATA COMES FROM

Structured
(inside the business)

- **POS —** What's selling, how much does it cost, who's buying it
- **Suppliers —** Product availability, prices
- **Accounting —** Costs, revenue, margins
- **Labor —** Wages, salaries, tips

Unstructured
(outside the business)

- **Social media —** Likes, trends, retweets, shares, comments
- **Customer profiles and loyalty programs —** Names, addresses, email, preferences
- **Weather and traffic patterns**

Why you need both

Structured data tells you the **"what"**; unstructured data tells you the **"why."** Using both gives you a more holistic view of your customer.

Figure 1: Image courtesy restaurant org - Data Sources

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

Big data analysis in Finance Sector

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

ABSTRACT

In order to understand what drives an organization or company profit, we want to 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.

KEYWORDS

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

1 INTRODUCTION

Big data as it's name implies, refers to large and complex data which continues to grow enormously day by day. There are huge number of sectors or applications including government, business, technology, universities, health-care, finance, manufacturing etc who make use of big data by obtaining meaningful information using big data technologies.[2] We investigates how big data is helpful in financial firms in terms of predictive analysis and profitable growth. The finance sector is generating huge amounts of data on a daily basis from products and marketing, banking, business, to share market. 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 a detailed view of customer segmentation in banking sector.[10]

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. [10]

Data storage: Even though big data does not defines 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.[7]

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.

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

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.

Data Lakes can be accessed by Machine learning algorithms, In-memory technologies, fast access DBs, big data queries and real-time analysis methods which consume less time to come up with meaningful information and reports.

Data lakes: Data Lakes can be compared to the actual lakes where water doesn't get filled like that instead there are rivers or streams that bring water to it. In data lakes this is called ingestion of data. we collect all the data that we required to analyze to reach our goal irrespective of the source. These fistreamsfi 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,fi), data from all sorts of logs (e.g weblogs, clickstream analysis,fi), XML, machine-to-machine, IoT and sensor data,, you name it (logs and XML are also called semi-structured data). There can be data filters in place based on the requirements.[4]

1.3 Identify business priorities

As per the six sigma methodology, the three steps for aligning projects to business priorities based on gross profits are,

* Identify the relative importance of strategic business objectives
* Identify the relative importance of specific key business processes
* Calculate the relative importance of key metrics of key processes

[5]There are different methodologies for identifying and prioritizing use cases for business priorities and innovations. The data-driven approach to find use cases includes BARC "Smart Data Science" methodology which apply techniques like pattern recognition and business intelligence. Industry communities are moving from traditional BI (data warehousing, latency, data sampling) to big data technologies

1.4 Customer Segmentation and personalized marketing

Banks have been under pressure to change from product-centric to customer-centric businesses. One way to achieve that transformation is to better understand their customers through segmentation. Big data enables them to group customers into distinct segments, which are defined by data sets that may include customer demographics, daily transactions, interactions with online and telephone customer service systems, and external data, such as the value of their homes. Promotions and marketing campaigns are then targeted to customers according to their segments.[1]

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

Personalized Marketing: One step beyond segment-based marketing is personalized marketing, which targets customers based on understanding of their individual buying habits. While itfis supported by big data analysis of merchant records, financial services firms can also incorporate unstructured data from their customers' social media profiles in order to create a fuller picture of the customers' needs through customer sentiment analysis. 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.[1]

1.5 Understand new business opportunities

Big data will fundamentally change the way businesses compete and operate. Companies that invest in and successfully derive value from their data will have a distinct advantage over their competitors fi!? a performance gap that will continue to grow as more relevant data is generated, emerging technologies and digital channels offer better acquisition and delivery mechanisms, and the technologies that enable faster, easier data analysis continue to develop. It is difficult to identify whats 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.[9]

Big data is the intersection of business strategy and data science, offering new opportunities to create competitive advantages. It allows companies to use data as a strategic asset, equipping them 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[10]

1.6 Discovery of innovation possibilities

Data is increasingly becoming a key differentiator between wildly profitable and struggling businesses. Exploring and analyzing data translates information into insight and drives to innovations. [3]

Firms are supposed to make decisions based on facts and data rather than intuition and should keep an open mind to innovation concepts.

1.7 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.[6]

Banks and financial services firms use analytics to differentiate fraudulent interactions from legitimate business transactions. By applying analytics and machine learning, they are able to define normal activity based on a customer's history and distinguish it from unusual behavior indicating fraud. The analysis systems suggest immediate actions, such as blocking irregular transactions, which stops fraud before it occurs and improves profitability.[1]

1.8 Risk Management

Financial firms especially banking sector are facing new regulatory requirements and challenges or risks each year. Big data adoption provide 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.[8]

1.9 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 fidata lakesfi can provide substantial cost advantages, due in part to greater scalability but also 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, CFOs can reduce the overall cost to serve in relation to data elaboration.[10]

Big Data adoption helps organizations simplify and reduce the costs of taking data from the source and converting it into useable information for regulatory reporting, including such data-intensive activities as real-time simulations and scenario analysis that are often required by the regulator.

2 BIG DATA - RISKS AND CONSIDERATIONS

Big data plays an increasingly important role in the financial services sector, where it is used for everything 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're developed and/or used in an improper way.[11]

Data Security: This risk is obvious and often uppermost in our minds when we are considering the logistics of data collection and analysis. Data theft is a rampant and growing area of crime fi?! and attacks are getting bigger and more damaging. [12]

Data Privacy: Closely related to the issue of security is privacy. But in addition to ensuring that peoplefis personal data are safe from criminals, you need to be sure that the sensitive information you are storing and collecting isnfit going to be divulged through less malevolent but equally damaging misuse by yourself or by people to whom you have delegated responsibility for analyzing and reporting on it.[12]

Bad Analytics: Aka figetting it wrong.fi Misinterpreting the patterns shown by your data and drawing causal links where there is in fact merely random coincidence is an obvious pitfall. Sales data may show a rise following a major sporting event, prompting you to draw a link between sports fans and your products or services, when in fact the rise is based on there being more people in town, and the rise would be equally dramatic after a large live music event.[12]

Bad Data: There might be situations where many data projects that start off on the wrong foot by collecting irrelevant, out of date, or erroneous data. This usually comes down to insufficient time being spent on designing the project strategy.[12]

3 CONCLUSION

The Big Data revolution, however, offers new opportunities for profitable growth, and financial services firms are responding enthusiastically. It has become their derived knowledge that making size-able investments in bigdata is ultimately a gain. Data has become the key element for decision making with the right choice of analytical tools and skillset. 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.

Apart from financial firms, big data continue to bring big benefits to our day to day life: advertisements focused on what you actually want to buy, smart cars that can help you avoid collisions, wearable devices that can monitor your health and notify your doctor if something is going wrong.

ACKNOWLEDGMENTS

Yet to add details here

REFERENCES

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

[2] [n. d.]. Big data. Web page. ([n. d.]). https://en.wikipedia.org/wiki/Big_data

[3] [n. d.]. Big data. Web page. ([n. d.]). 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

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

[5] [n. d.]. Three Steps for Aligning Projects to Business Priorities. Web page. ([n. d.]). https://www.isixsigma.com/methodology/metrics/three-steps-aligning-projects-business-priorities/

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

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

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

[9] 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

[10] Gregor Meyer Fabrizio Sarrocco, Vincenzo Morabito. 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

[11] 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

[12] Bernard Marr. 2015. The 5 Biggest Risks of Big Data. (June 2015). http://data-informed.com/the-5-biggest-risks-of-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. One solution of this problem is to do some processing on the edge devices. This is extremely helpful in providing responsive and real time analytics.

KEYWORDS

hid201, i523, Edge Computing, Big Data Analytics

1 INTRODUCTION

Internet of things is rapidly gaining importance and ”according to the Evans Data Corporationfis Global Developer Population and Demographics Study, some 6.2 million developers are already working on IoT applications.”[10] With the rapid increase in the acceptance of Internet of Things (IoT) devices across various fields in 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.[12]. 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 perfomed or device-dentric approach, where devices have some proprietery code and perform analytics in a stand-alone manner”[12]. Networks are largely centralized with organizations soring all data, which may not be directly beneficial for them, in their data centers, and data flowing from the edge to the cloud on each operation[1]

With increase in the number of connected devices, it gets increasingly dificult 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 edge[13][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 cost of computation and storing data on the cloud[12].

2 HOW EDGE COMPUTING WORKS

Edge computing emerged with the development of content delivery networks (CDNs) by Akami which use nodes closer to the user to prefetch and cache web content to accelerate web throughput.Edge computing extends this concept with the help of cloud infrastructure to run arbitrary task specific code at 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.[11]

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 virtual reality applications which typically require a latency of less than 16ms to appear stable.[2][11] Proximity also increases scalability with the help of edge analytics which uses the cloudlet to perform 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 devices increases.[11]. 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]

In industrial applications like aviation where a large amount of data is generated on each flight[11], analyzing this data in a centralized manner becomes impractical. In such cases fog computing is more useful which adds a ”hierarchy of elements 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 proprietery 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 ”that classifies vehicles from video streams using a Tensorflow deep neural network encapsulated within a NiFi dataflow executing across multiple Pis. This helps with local analytics of video data streams close to the camera source, but with the flexibility of using the same deployment in the Cloud too, say, when the edge is constrained.”[12]

Yang Zhao et. al proposed an occupancy and activity monitoring appicattion 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 dvelopers provide a web portal to help ease monitoring activity from a remote location.[15]

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 specially if the cameras are connected wirelesly. In addition to this all of the 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 surveiillance 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.[11]

A real time video processing solution is proposed in [2] that focuses on traffic plannin and safety and provide high accuracy outputs and detects anomalous traffic patternd to suggest prermptive 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 sppacial knowledge, and must rely on cloudlet based edge solution to provide seemless interaction for the users.[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 would 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 NEW APPROACHES

[9] [4]

5 AI ON THE EDGE

With the emergence of decentralized aplications, smart mahines that rely on machilne learning and mesh computing to provide local real time analytics are becoming a reality. MIT's Eyeris 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 intelligenc (AI) applications need a huge amount of processing power and rewuire a large amount of data, traditional AI applications rely on cloud servers to perform their computation. This is a serious limitation in applicaitons where connectivity is not reliable and time-critical decisions are required. [5]iEx.ec is a company that uses Etherium blockchain to createa market for computing resources, facilitating distributed mahcine 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 leraning algotithms can offer intelligence at the edge devices. Movidius recently announced a deep learning compute stick that can add achine learning capabilities to computers and raspberry pis as a plug and play device.[5]

Machine learning algorithms line one-shot learning that require less data are rapidly enabling edge devices to perform intelligent tasks easily.[14]. Gamalon, backed by Defense Advanced Research Projects Agency (DARPA), is using Bayesean Program Synthesis to reduce the amount of data required for machine learning.[5]

6 CONCLUSION

With the increase in the number of conected devices and the increase in the demand of real time and interactive applcations, edge computing is a necessity and many industries are rapidly moving towards edge solutions. Although industrial IoT is gaining importance but still faces challenges with the integration of legacy applications and proprietary applications with new technology, new open source solutions are gaining importance. With the emergence of decentrallized applications and the growing importance of machine learning, edge computing is required as a foundation to move towards decentralized AI applications, that 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 whih 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] 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

[10] 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

[11] Mahadev Satyanarayanan. 2017. The Emergence of Edge Computing. *Computer, IEEE compter society* (2017). http://elijah.cs.cmu.edu/DOCS/satya-edge2016.pdf

[12] 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

[13] 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].

[14] 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].

[15] 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

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Rono-Hills
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
30 Shuangqing Rd
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

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 with determining 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. The purpose is to explore at a very high level how higher education could utilize big data and 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. 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 [3]. 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" [3]. Higher education could benefit from using big data in their marketing efforts for recruiting and enrolling students as well as identify what gaps may still exist in the quest to understand todayfis college student in their search on which university to attend.

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 [2]. Meanwhile, the number of four-year institutions in the United States has increased with more than 3,000 available college options [4]. 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 [1] identified a number of aspects that will impact the higher education landscape, among those included are:

- 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 (CITE SOURCE).

- The college search process today happens primarily in the digital space which includes third-party websites, email, social media, and digital advertising. 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.

- 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 and analytics can be used to help assess the impact of these trends as well as how institutions can use data to inform decision making.

3 BIG DATA TO SEGMENT STUDENTS 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. 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 TO UNDERSTAND STUDENT 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 state that the institution's website is their most used resource when exploring colleges [5]. 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 [7]. 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 TO CONVEY VALUE

Utilizing big data to understand outcomes of current students, and ultimately graduates, can help tell the value story to prospective students. By tracking the experiences among currently enrolled 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 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 in higher education marketing to be successful, there are basic measures that have to be met. Marsh et. al outlines some key considerations when using big data for effective decision-making which include: accessibility, quality, timeliness, and motivation to use [6]. These same factors can be applied to utilizing big data in higher education marketing.

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 [8]. As Daniels state, there is also data in higher education that lives across a number of areas in a wide variety of formats [3]. With this, accessing data can be a challenge as their is no central system or warehouse. Depending on the type of data needed, the sources can be siloed which means that the data sources are not connected to one another to provide a complete picture. Further, the level of permissions 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. Insights around data are only as good as the people that make use of them. 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 [6]. 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.

6.4 Motivation

There is also a underlying cultural aspect to using big data analytics in the right way across an institution. With the silos that can 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) [3]. The importance and message about utilizing big data analytics has to come from leadership.

7 POSSIBLE SOLUTIONS

While significant challenges can exist in utilizing big data 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 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

[Figure 1 about here.]

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 place 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 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 this information should be stored.

9 CONCLUSIONS

Competition for today's student will only increase with changing educational needs and offerings, including development of emerging degree programs as well as delivery, including online classes. 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. Further, there has to be a culture within an institution that motivates others to make data-driven decisions.

REFERENCES

[1] 2014. Trends in Higher Education Marketing, Recruiting, and Technology. (2014).

[2] Peace Bransberger. 2017. Impact and Implications: Projections of Male & Female High School Graduates. (2017).

[3] Ben Daniel. 2015. Big Data and Analytics in Higher Education: Opportunities and Challenges. *British Educational Research Association* (2015).

[4] National Center for Education Statistics. 2015. Digest of Educations Statistics. (2015). https://nces.ed.gov/fastfacts/display.asp?id=84

[5] Stephanie Geyer. 2016. E-Expectations Trend Report. (2016).

[6] Julie A. Marsh, John F. Pane, and Lara S. Hamilton. 2006. Making Sense of Data-Driven Decision Making in Education. (2006). www.rand.org

[7] Mohammad Amin Omidvar, Vahid Reza Mirabi, and Narjes Shokry. 2011. Analyzing the Impact of Visitors on Page Views with Google Analytics. 2 (2011).

[8] 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* (2012).

[9] Jeffrey Selingo (Ed.). 2017. How Colleges Use Big Data to Target the Students They Want. (2017).

[10] Sharon Slade and Paul Prinsloo. 2013. Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist* (2013).

[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).

List of Figures

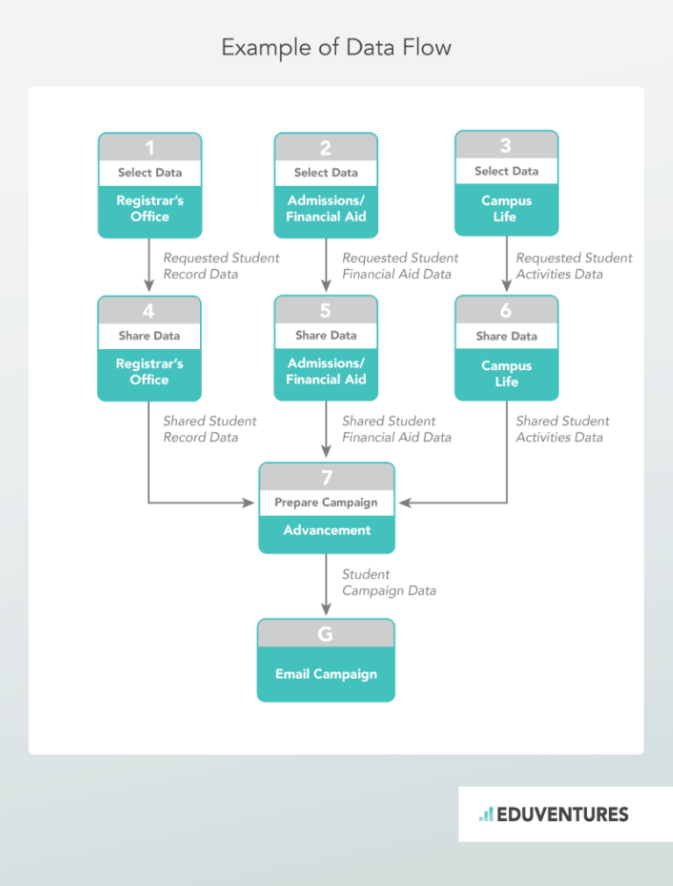

Example of Data Flow
1
Select Data
Registrar's Office
2
Select Data
Admissions/ Financial Aid
3
Select Data
Campus Life
Requested Student Record Data
Requested Student Financial Aid Data
Requested Student Activities Data
4
Share Data
Registrar's Office
5
Share Data
Admissions/ Financial Aid
6
Share Data
Campus Life
Shared Student Record Data
Shared Student Financial Aid Data
Shared Student Activities Data
7
Prepare Campaign
Advancement
Student Campaign Data
G
Email Campaign
EDUVENTURES

Figure 1: Example of a data flow [11]

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. This paper discusses about the implementation of Big Data technologies, challenges of implementing Big Data in Electric Power Distribution Systems. [1]

KEYWORDS

Big Data, Power Distribution,Smart Power

1 INTRODUCTION

Volume of data is increasing. According to forbes, it is said that, worldfis 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 Data Sources

Smart meters which are placed at customerfis vicinity will record the consumption of a specific group of customersfi. 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.

Power 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.

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

1.2 4 v’s in Big Data in Power Distribution System

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.

REFERENCES

[1] Amr A. Munshi and Yasser A.-R.I.Mohamed. 2017. Big data framework for analytics in smart grid. *Electric Power Systems* (2017).

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

Big Data Analytics in Agriculture

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

ABSTRACT

Big Data and Data Science is significantly impacting the industry of agriculture and food safety in the food supply chain. 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]. The availability the wireless technology, the Internet of Things, smart machines, and sensors are making food production processes increasingly data driven and data enabled [7]. This in turn, is making farming production processes more productive and food delivery systems 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]. Large amounts of information is being captured, analyzed, and used to make operational decisions [4]. 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 [6]. 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 [3]. 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 [4].

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 [3]. 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 [5].

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 [2]. 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 [1], 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 [5], 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 [4], 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. [2]. 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 [1].

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 [4]. For example, soil characteristics such as texture, organic matter, and fertility is collected to the meter at locations throughout the world [4].

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 [5]. Now, in addition to selling seeds and machinery, these companies are selling decision support and data science services [3].

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 [5].

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) [5].

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 [2]. 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 [6]. In this way, disease outbreaks can be recognized and addressed immediately [5]. 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 [5]. This same technology can even tell if a single seed has been missed [3]. GPS units on tractors, combines, and trucks help determine the optimal usage of equipment [6].

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. [4]. Sensors indicate when livestock are ready to inseminate or give birth [1]. 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 [3]. 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 [4]. 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 [4].

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 [6]. 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 [6].

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 [4].

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.[4]. 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.[2].

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 [5]. 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? [3]. 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

Improvements to agricultural productivity as result of big data technology are beyond substantial. Big data is being referred to as the most significant revolution in farming productivity since mechanization. In 2009, the United Nations estimated that 900 people in the world were undernourished and that 65 countries face alarming food shortages [4]. Big Data is expected to make an impact Food Insecurity throughout the world as farmers throughout the world adopt these techniques. This technology will enable even small holder farmers to make full use of their productive potential. The use of precision farming techniques and digital technologies will enable farmers to maximize the use of every inch of soil and even the production of each individual plant.

Big Data is improving the food delivery system. Information is available to producers and suppliers that in the past has been impossible to obtain [7]. Big data is making the food supply healthier and safer. Big Data in Agriculture is here to stay.

REFERENCES

[1] 2017. Farming goes digital - The 3rd Green Revolution. Web page. (Sept. 2017). https://www.cema-agri.org/page/farming-goes-digital-3rd-green-revolution

[2] 2017. Precision Farming. Web page. (Sept. 2017). https://en.wikipedia.org/wiki/Precisionagriculture

[3] Dan Bobloff. 2017. Big Data Comes to the Farm. Web page. (Sept. 2017). https://www.businessinsider.com/big-data-and-farming-2015-8/

[4] Nir Kshetri. 2016. *Big Data's Big Potential in Developing Economies.* CABI.

[5] Katherine Noyes. 2017. Cropping Up on Every Farm Big Data Technology. Web page. (Sept. 2017). https://www.fortune.com/2014/05/30/cropping-up-on-every-farm-big-data-technology

[6] Sparapani. 2017. How Big Data and Tech Will Improve Agriculture from Farm to Table. Web page. (Sept. 2017). https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-tech-will-improve-agriculure-farm-to-table/

[7] Sjaak Wolfert. 2017. Big Data in Smart Farming - A review. *Agricultural Systems* 153 (Feb. 2017), 69–80. http; /sciencedirect.com/science/article/pii/s0308521x16303754

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

As waste management becomes a greater concern for cities and municipalities around the world with increase in population and the 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 government, companies, etc.

KEYWORDS

Waste Management, Big Data, Local Government

1 INTRODUCTION

In the current fast paced society, as production of goods increases and new distribution chains constantly change, the production of disposed materials and goods, from now on called solid waste, has increased over the past ten years, going from around 0.64 kg per person per day of solid waste to approximately 1.2 kg per person per day, and it is expected to increase to about 1.42 kg by 2025. [13] This causes the problem of waste management to increase in complexity and magnitude.

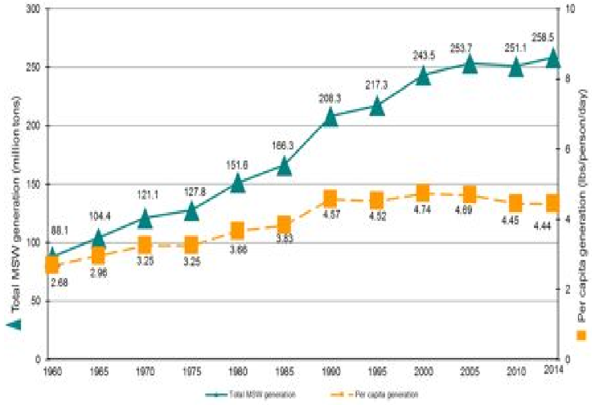
Because of this, different local governments and organizations have seen the need to develop regulations to control the different features, segments, processes of the action of disposal from the moment the material is discarded until the moment the material reaches it's ultimate destination like recycling plant or landfill. This set of systematic regulations is called solid waste management. Simple techniques are used in earlier days to make decisions for waste management which is to choose an option from multiple available options [1]. Decision-making became much more complex when multiple parameters adds up to the system.

Classifications of solid waste is determined by its sources, various types of wastes accumulated and the rate of disposal are to be constantly monitored and controlled in parallel along with improving current systems [5]. Large volumes of data will be collected on daily basis from each classification of solid waste which includes multiple parameters. Multivariate data analysis methods [4] provides an exploratory data analysis, classification and parameter prediction using this data.

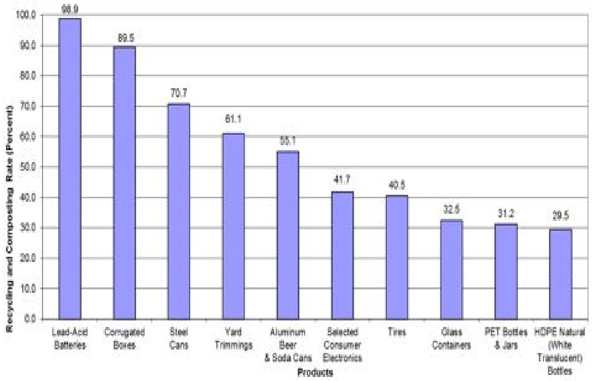
2 MUNICIPAL SOLID WASTE MANAGEMENT

Municipal Solid Waste (MSW) commonly termed as the garbage or trash consists of items we use in everyday life like food leftovers, plastic bottles, wooden furniture, electrical and electronic appliances, glass, medical waste, cardboards, waste tires, office wastes, consumer goods etc which comes from residential, commercial, institutional and industrial sites.

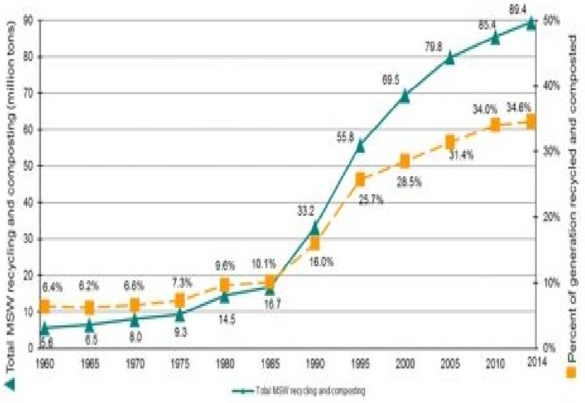
The amounts of disposed material and it's composition vary depending on the country, place and activity that is performed at the site where the waste is generated [5]. According to EPA statistics 2014, Americans have generated about 258 million tons of MSW in which more than 89 million tons is recycled and composted. This is equivalent to 34.6% recycling rate compared to 6.4% in 1960. In addition, 33 million tons of trash is combusted for energy and 136 million tons were landfilled. Figures below represents MSW generation rates, recycling, composition rates and Total MSW generation between 1960 and 2014 [7].



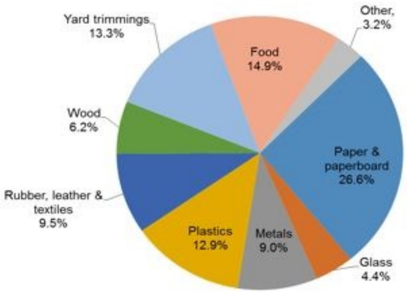
MSW Generation Rates 1960 to 2014



Recycling and Composting Rates of Selected Products, 2014 (does not include combustion with energy recovery



MSW Recycling and Composting Rates, 1960 to 2014



Total MSW Generation (by material), 2014
258 Million Tons (before recycling, composting, or combustion with energy recovery)

There are also important differences between the general composition of the waste generated in rural area and what is produced in urban area, the waste produced in the later is highly influenced by the culture and the practices of our modern society. [5]

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 each case.

In general, decisions can be classified as optimal, good, or fortuitous. [1] and this can be applied to Waste Management.

Having that Good decision-making is mostly based on experience, comparison of elements and trial and error, and that fortuitous decision-making have no scientific base; one must always try to solve the problem -in this case waste management related- with Optimal Decision making, that requires techniques and technologies provided by other fields. [1]

3 BIG DATA AND WASTE MANAGEMENT

By collecting and storing large volumes of data related to types of waste, quantities, periodicity, and composition; usually from independent sources. Big data can be interpreted in a way that allows the different actors that intervene in Waste Management to make Optimal Decisions. [16] Big Data refers to taking very large amount of data sets and applying technologies to analyze these data sets. As stated in "Fourth Paradigm" [12], big data exploration is about finding patterns in data, analyzing the trends and causalities.

Big Data can be used for strategic policy making in almost any field and the Greater Manchester Waste Disposal Authority (GMWDA), England's largest Waste Disposal Authority, has turned Big Data to better plan their services. In order to achieve that, they are collaborating with the University of Manchester who uses the data generated by the GMWDA. Together they help create environmentally sustainable solutions for Manchester and the 1.1 million tonnes of waste that is produced each year [15].

Big data will help governments to track the amount of disposals at different locations and their quantity in-order to generate heat maps of locations with largest waste collected and will help to improve necessary solutions for better environments [15]. Waste Management is not only government issue. Citizens should take initiative and educate others on how to recycle waste for their better living. With the help of collected data, governments will notify citizens about the importance of waste management through mobile phones as its considered the cheapest means of communication in modern world.

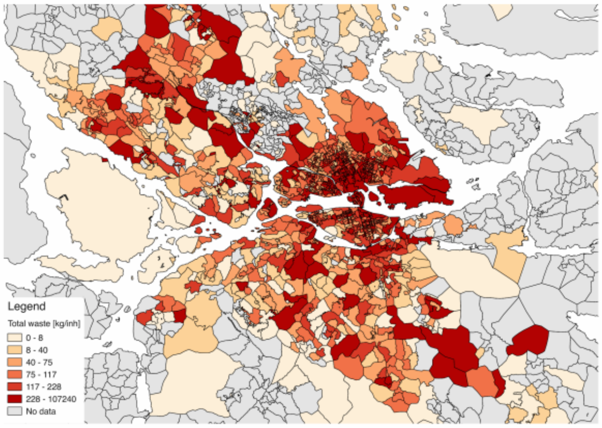
3.1 Solutions for effective Waste Management

Purpose of Big data in waste management is to facilitate municipal government bodies on how much waste is disposed, environmental pollution, rate at which waste is recycled, optimizing routes to reduce the time and money.

One such solution is being implemented in an upcoming smart city of Songdo, a chip card is made mandatory for every citizen to use while disposing their garbage. Data collected from these chip cards will be used for analyzing on the quantity of waste disposed, and their locations. Each trash bin is incorporated with sensors to provide height of the garbage accumulated, temperature and air pollution levels. These multiple parameters help municipal authorities to forecast perfect timings to collect the trash and optimize the routes to save time and their cost [15].

Researchers in Ethiopia are combining socioeconomic data along with geographic data to get a clear understanding on the patterns of how household waste is being properly distributed and collected. This study helped local authorities to better manage waste practices in urban areas [15].

A group of researcher from University of Stockholm are using Big Data to identify on how the transportation of waste and the garbage collection routes can be optimized in the city. Using wide variety of data sets collected from various sources, roughly around half a million entries of waste fractions(as showin in fig below), trash bin locations, weights, truck routes researchers have develop waste generation maps of Stockholm. This research help to reveal quite a few inefficiencies the local government is facing and help to improve the waste management [15].



Choropleth of all waste in the dataset, aggregated per zip code and normalized to population.

3.1.1 Vehicle Routing Problem. Vehicle route optimization is one of the main concern in waste management. It is generally termed as Vehicle Routing Problem(VRP) [6]. Given a common problem to a general heuristic a strong solutions can be modeled manually to solve it. But in a real-world multiple factors will be influenced either directly or indirectly to that problem. Common known factors that shows influence on vehicle routing problems are number of vehicles, garbage collection stops and the route length. Depending on the complexity of problem, few more factors can be included like vehicle types and disposal facilities.

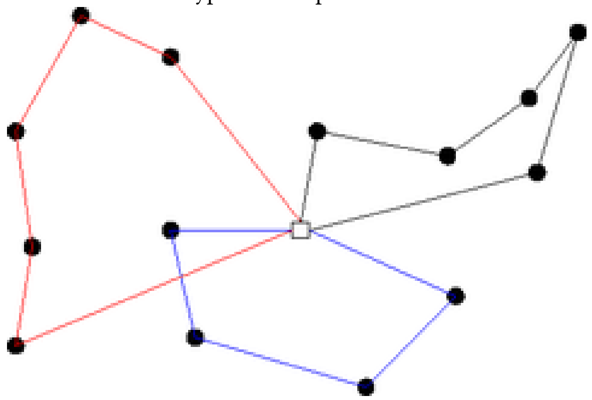


Figure represents Vehicle Routing Problem

Two of the most basic VRPs are the Travelling Salesman Problem (TSP) and the Chinese Postman Problem (CPP) according to Joroen, Liesje and Jonas [2]. But when too many constraints and attributes are considered, both of the TSP and CPP tends to get harder to solve problems. Many researchers had made various publications since 1995 on waste management vehicle routing problems as shown in Figure 1 and yet the problem still persists. Mathematical models need to be developed to provide city administrators with a tool to make effective long-and short-term decisions relating to their municipal disposal system [3].

In modern world, not only direct impacted attributes causing VRP problems. But indirect attributes like daily traffic, weather conditions, energy prices, demand fluctuations, vehicle health, dump site inventory also affecting to strengthen the worst. Research team at OSI came up with a better solution for solving VRP problems using Big Data technologies. Mixed Integer Programming (MIP) formulation interacts with millions of attributes in a live environments providing real-time decisions to optimize the VRP. Big Data technologies are used to enable prediction of travel times, address demand forecasting on a tactical time horizon. This approach showed a tremendous improvement in forecasting part of the VRP problem at a range of 5% to 10% [10]. Any improvement, even less than 5% created on VRP is a significant improvement [11].

In 2011, Faccio, Persona and Zanin [8] investigated the feasibility of communication between bins, collection vehicles and a central operator. The waste bins can be fitted with a volumetric sensor, RFID and GPRS communication and can send information about their status. Using this real-time data, routes can be optimized in order to make optimal use of the vehicle's cargo space. Waste containers that still have not reached a certain threshold to be emptied are skipped by, saving valuable travel time and distance. Also of importance, fewer waste collection vehicles were needed. A key finding was that the economic feasibility of providing a sensor network to support waste management in this case, was estimated to a payback period of roughly three years [14].

Real-time, on-demand routing is helping now to address operating costs and service improvements. Trash collection vehicles are closely monitored with a remote self-diagnostics to identify vehicle health, required repairs and pre-ordering of replacement parts. This will help to prevent downtime of garbage collection trucks from being getting repairs. Hand-held applications and tools for autonomous service verification are being used in measuring program success and to outreach programs [9].

3.1.2 Disposal of Landfills Problem. The process of solving a math program requires a large number of calculations and is, therefore, best performed by a computer program. [1]

4 OPPORTUNITIES FOR WASTE MANAGEMENT OPTIMIZATION

4.1 Statistics and Waste Management

There are many data analysis methods that are used when studying waste management, but the two most popular are PCA and PLS1. [4]

Lingo is a mathematical modeling language designed particularly for formulating and solving a wide variety of optimization problems including linear programing. Lingo optimization software uses branch and bound methods to solve problems of this type. [1]

4.2 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, there are solutions that can help interpret large data sets, apply statistics and algorithms of different kinds and display them in a way that make reference to a geographical space.

The second category of studies focuses on minimizing transportation of waste collection through optimal routing algorithms. For example Kim et al [18] use two methods to calculate an optimal set of routes, the first being Solomon's insertion algorithm, the second being a clustering algorithm. Their aim was to minimize the driven distance, as well as to balance the workload. At the same time, the constraint of legally prescribed lunch breaks (so called time-window problem) had to be satisfied. McLeod and Cherrett [19] suggested a route optimization for three areas and connected waste companies in North Hampshire (UK). By applying simple rerouting, sharing of routes between the 3 areas and adding vehicle depots at the waste disposal sites, they estimated annual savings as large as 10,000 km for the studied routes (this covers one fifth of all routes in North Hampshire).

Another study performed by Wy, Kim and Kim [20] studied a routing algorithm for waste collection using roll-on/roll-off containers, again while factoring in the time windows. Buhrkal, Larsen and Ropke [21] were one of the ffirst to suggest the environmental importance of optimizing waste collection itineraries. They utilized an adaptive large neighborhood search algorithm, and a clustering method and their scope was residential waste collection. Depending on the computation time, using the actual collection points and lunch time windows, the savings amounted to 13 percent average. With larger time windows and better starting conditions, heuristics with a distance reduction of up to 45% could be achieved [14].

Many data analysis methods are used when studying waste management, but the two most popular are PCA and PLS1. [4]

5 CONCLUSIONS

There are different tools to optimize the different waste management practices and to improve the information available for decision makers. Local governments had just started to adopt Big Data technologies for solving problems involved in MSW. In future using Big data Analytics, large amounts of data sets will be used to identify trends and patterns that could highlight improvement opportunities. Big Data will play a major lead role in managing better smart cities and government authorities will be benefited with tremendous improvements in waste management. Thanks to Big Data Analytics for making smart cities much more effective and efficient.

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] 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

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

[7] 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

[8] 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

[9] 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

[10] 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/

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

[12] 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

[13] 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

[14] 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.

[15] 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

[16] 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.

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 47405
mkagita@iu.edu

ABSTRACT

As waste management becomes a greater concern for cities and municipalities around the world, 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 government, companies, etc.

KEYWORDS

Waste Management, Big Data, Local Government

1 INTRODUCTION

Concept of waste managementfi

Solid Waste Management (SWM) is a set of consistent and systematic regulations related to control generation, storage, collection, transportation, processing and land filling of wastes according to the best public health principles, economy, preservation of resources, aesthetics, other environmental requirements and what the public attends to [1]

Managing solid waste is one of the most essential services which often fails due to rapid urbanization along with changes in the waste quantity and composition. Quantity and composition vary from country to country making them difficult to adopt for waste management system which may be successful at other places. Quantity and composition of solid waste vary from place to place [3]

2 OPPORTUNITIES FOR WASTE MANAGEMENT OPTIMIZATION

By collecting and storing data related to types of waste, quantities, periodicity and composition.

2.1 GIS Analytics

3 STATISTICS AND WASTE MANAGEMENT

While rural area usually generates organic and biodegradable, urban area produces waste influenced by culture and practices of society. [3] p47 to 63

There are many data analysis methods that are used when studying waste management, but the two most popular are PCA and PLS1. [2]

decision makers should distinguish between optimal, good, and fortuitous decision-making. In the optimal decision making, one can solve the optimal problem using the techniques available in other fields. In this solution method, generally some constraints (criteria) are consid- ered, where the function(s) is to be optimized through applying some methods. Good decision-making is done based on experience, trial and error or comparison between different options of the integrated SWM. Although it is possible to choose decisions close to the optimal state using this decision-making method, today these methods are not applicable due to increased number of different combinations in the decision-making process. In the fortuitous decision-making, since decisions are made with no scientific base, so the results are not acceptable [1]

The process of solving a math program requires a large number of calculations and is, therefore, best performed by a computer program. Lingo is a mathematical model- ing language designed particularly for formulating and solving a wide variety of optimization problems including linear programing. Lingo optimization software uses branch and bound methods to solve problems of this type. [1]

4 CONCLUSIONS

Working on this

Generated by bibtex from your `.bib` file. Run latex, then bibtex, then latex twice (to resolve references) to create the `.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

A MORE HELP FOR THE HARDY

Of course, reading the source code is always useful. The file `acmart.pdf` contains both the user guide and the commented code.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

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] 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, 24. https://doi.org/10.5772/53975

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

Automated Information Extraction in Electronic Health Records

Nicholas J Hotz
Indiana University
nhotz@iu.edu

ABSTRACT

Electronic medical records (EMRs) play an increasingly important role in health care. However, the rapidly growing volume of text in EMRs creates challenges in the extraction of information. As such, many research institutions are developing computer-based systems to automate EMR structured information extraction (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 comprehensively 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. However, recall seems sufficient for commercial recommendation systems to support manual coders and for audit purposes.

KEYWORDS

Natural Language Processing, Information Extraction, Clinical Coding, Electronic Health Records

1 INTRODUCTION

Demand for structured health data continues to grow [16], 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 [11][14]. However, EMRs also create challenges, in part because EMR information is often stored in narrative form which describe patients, their own and their family's medical history, their personal lifestyle, and their current medical conditions. [11] 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 [16][11].

Commonly studied clinical NLP problems include de-identification [19], the development of patient problem summaries [6], and diagnostic code extraction [12]. 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 that 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” [12]. According to Kavauluru et al, leaf nodes are codes that provide specific information used for ”billing and reimbursement, quality control, epidemiological studies, and cohort identification for clinical trials” [9].

Currently, coding professionals and physicians 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 [16] [1] [14][19]. This complexity limits manual coding accuracy. Manual coders often disagree [13] and are more specific than sensitive in their code assignments [3]. Errors are prevalent; for example a Sweedish study of 4,200 patient records found errors in 20% of the main diagnoses [19]. Over-coding can be viewed as fraudulent because health care providers would bill for services not rendered while under-coding prevents providers from earning reimbursements for valid conditions and services [12].

Since the 1990s [8], researchers have tried to improve the coding processes through automated coding and classification technologies which, according to Stanfill et al, ”encompass 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”[16]. In 2004, the American Health Information Management Association asserted that ”The industry needs automated solutions to allow the coding process to become more productive, efficient, accurate, and consistentfi [16]. However, Stanfill et al. conclude in their 2010 literature review that the relative performance of automated systems to manual coding is not yet known[16]. As of 2008 and still in 2015, automated systems are still mostly used for research purposes with few applications in use by practitioners [11][19].

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 [11]. 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 [2]. 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 Information Portability and Protection Act (HIPAA) and the European Union's General Data Protection Regulation (GDPR); consequently large corpora are rarely available from other health systems [11][16].

As a related problem, even when corpora are available, the annotation process is time-consuming, expensive, and traditionally relies on domain experts and linguists [11][19]. 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 are 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 is around 10 % [15], which is higher than most texts and complicates several 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 [11]. 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 literate 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 clincal text, leveraging partially-automated processes to facilitate the annotation of corpora, and focusing on multiple languages [5] [19].

3 EMR INFORMATION EXTRACTION 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 for medical texts including:

- **Language Detection:** Multi-lingual studies may start with language detection algorithms, although some might still rely on manual detection [6].
- **Spell checking:** Clinical NLP spell checking uses standard dictionaries and medical-specific tools such as unified medical language system (UMLS) and WordNet [11].
- **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) [11].
- **Tokenization and sentence-splitting:** Tomanek et al. find that the training corpus is not too important for sentence-splitting but is crucial for tokenization, the process for breaking text into tokens such as words, phrases, or symbols [17][6].
- **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 [11][6].
- **Parsers:** Parsers identify the sentence syntax, word dependencies, and expressions of interest [11][6].

Context feature detection and analysis typically follows the above steps and identifies how words and concepts are being in the context of the sentence. Clinical NLP systems often leverage 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 currently has the condition or if the text references their medical history), subject identification (e.g. to identify if the condition belongs to the patient or some one else such as a family member), and severity (such as mild, moderate, or severe conditions) [11][19].

4 EFFECTIVENESS OF AUTOMATED ICD9 CODE EXTRACTION OVERVIEW

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 two or more independent reviewers with an adjudication process for disagreements, and 51% use the regular practice standard of one reviewer [16]. Although considered more reliable, gold standards are still prone to error [12]. The most commonly reported metrics include recall or sensitivity (69%), PPV or precision (46%), specificity (43%), and accuracy (25%) [16].

Most studies foucs only on a specific subset of clinical texts or diagnoses such as subdomains like radiology [14], for specific diagnoses like congestive heart failure [7] or cancer [10], or to extract only attributes of patients like smoking status [18]. 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 [16].

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, they achieved an F-measure of 39.5% with a 30.0% recall and 57.7% precision. They also attempted a flat SVM which returned an 27.6% F-measure with 16.4% recall but with a higher precision (86.7%) [12].

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 sumdomains such as radiology or pathology, chain classifiers perform best because codes are highly related 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 [9].

5 OUTLOOK

Add in stuff. No citations allowed. Argue for more practical-focused work.

REFERENCES

[1] 2013. Automated Coding Workflow and CAC Practice Guidance (2013 update). (11 2013). http://bok.ahima.org/PB/CACGuidance#.WchAZMiGOUl

[2] 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.

[3] 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.

[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] 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, S–84.

[6] 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, 182–186.

[7] 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. In *AMIA annual symposium proceedings*, Vol. 2006. American Medical Informatics Association, 269.

[8] Ramakanth Kavuluru, Sifei Han, and Daniel Harris. 2013. Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques. In *Canadian Conference on Artificial Intelligence*. Springer, 77–88.

[9] 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.

[10] 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, 420.

[11] 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.

[12] 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.

[13] 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, 97–104.

[14] 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.

[15] 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.

[16] 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.

[17] 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*. 49–57.

[18] 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.

[19] 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.

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [**?**].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

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

Mental Health Treatment

1 INTRODUCTION

Big Idea: Mental Illness is a big societal problem, which could benefit from a big data solution.

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 takes this processed data, and creates 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 [?].

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 [?]. A 2014 meta-analysis study estimated that the worldwide prevalence of mental illness was 17.6% and that 29.2% of the world population would experience mental illness at some point during their life [?]. 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 [?]. 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 [?]. It is estimated that mental health issues had a $100 billion cost on the US economy in 2002 [?], and in 2015 there were over 12,000 mental health treatment facilities in the US [?].

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 is still follows a fairly consistent pattern. First the mental health issue is identified [?], then treatment interventions are assigned [?], and finally treatment progress is monitored [?].

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 [?]. 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 strategies to change unhelpful thoughts, feelings, or actions, medication to reduce symptoms of mental illness, and supportive services such as case management to help coordinate efforts towards the person's goals [?]. Treatment monitoring is essential to the treatment life-cycle, 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 [?].

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, which explored using social media to identify mental illness in the general population, could demonstrate such an improvement. 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. [?].

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. [?]. 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 [?].

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. [?] Predictive risk profiles for patient's with bipolar 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 [?]. Predictive models show promise of being an effective big data application in mental health treatment, but require further advances in machine learning techniques and validated on larger samples before they can be widely administered [?].

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 [?]. 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 [?], a technique called a recommender system [?]. Also, as interactive interfaces are developed and used by large numbers of online users [?], 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. Utilizing text message or application notifications, treatment goal reminders, symptom assessment questions, or encouraging messages are sent to the treatment participant [?]. 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 [?].

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 can be done using clinically informed algorithms or machine learning paired with self-report [?]. Devices used were not just smartphones, but including wearables and an 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 [?].

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 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 [?]. 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 [?]. When mental health privacy is breached, discrimination regarding employment, insurance, housing, etc. are possible [?]. 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 [?].

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 [?]. It is also unclear who owns the data in these situations, causing more disruption [?]. 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 [?].

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 "data-rich" science [?]. 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 [?]. The integration of private commercial data for big data analytics is also a goal of some researchers [?]. 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 [?]. This is a point where public policy has fallen behind technological innovation, and that an inter-disciplinary effort from legal, data science, and mental health experts may be required to strike the balance between science and citizen security [?].

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 is generating considerable attention, but is 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. Even more nascent is integrating big data analytics web-based mental health interventions. 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 have 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

REFERENCES

Big Data Applications and Analysis in Maternal Death During Childbirth in 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 doubled from 2000-2014 and more than half of such incidents could be prevented with the current medical technology.

Most of the cases were result of medical error and unprepared hospitals. Doctorfis ability to protect the health of mothers in childbirth is a basic measure of a societyfis 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 [15] and [10].

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 American 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 [12].

[Figure 1 about here.]

From 1990 to 2014 pregnancy related death increased by 1.7 percent while worldwide that rate decreased by 1.3 percent. 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 [11].

[Figure 2 about here.]

Women giving birth in Asia have lower risk to die than those giving birth in United States [15].

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.

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 [9] and [4].

Additionally, Pregnancy Mortality Surveillance System was implemented in 1896, because of limited pregnancy death related records [8]. 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 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 [14]. Deaths related to pregnancy from 1998-2010 were published in Obstetrics and Gynecology journal [16]. 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 [14]. Figure 3 shows changes in pregnancy related mortality ratio in United States from 1987-2011 [7].

[Figure 3 about here.]

2 BIG DATA USAGE AND HOW IT CAN HELP

The causes of these death are not yet identified since only limited amount of data was analyzed [5].

Big data tools help to understand and organize pregnancy related deaths and causes. Also it helps to collect and identify risks by race ethnicity, economical status and age. Further examination of structured and unstructured data could help with preventing causes of pregnancy related death.

A similar study was done on October 8, 2016 by journal The Lancet, that called ”Global, regional, and national levels of maternal mortality, 1990fi?!2015: a systematic analysis for the Global Burden ofDisease Study 2015” [11]. They used a standardized process to identify, extract and process all relevant data sources. Standardized algorithms were implemented to adjust for age-specific, year-specific, and geography-specific patterns of incompleteness, as well as patterns of miss-classification of deaths [13].

Internet Of Things could be used to monitor patients and their pregnancy risks such as diabetes level or blood pressure. It could also track prescribed medicine, itfis especially useful for patients without health insurances [11].

Predictive analytic should be used, womenfis information could be shared between doctors and hospitals to be diagnosed in advance, improving number of healthy pregnancies. By being able to analyze big data, pregnancy risks will be predicted and provide women with safety and better pregnancy outcomes. The more analyzed data we have, the sooner it will reduce the mortality rates and it will be easier to diagnose each case. Special kits with appropriate medicine could be supplied to each hospital for individual patient.

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

The data could be put into Hadoop to make a more scaleable analysis with that. Itfis one of the most popular data management option. As of today, it’s one of the largest systems that is being used by many companies. Its ability to handle wide amount of data makes it efficient and provides possibility to get more accurate causes and reasons of maternity deaths. 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 to process huge amount of unstructured data and print out analyzed information. This system is all about handing the Big Data [6].

3 CONCLUSION

Pregnancy-related mortality findings should be recorded and cross analyzed. It provides a better view, results clarification and better health management.

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

All these years, there was not enough information that was structured for deeper analysis. Big Data getting larger daily, useful information is everywhere around us; including emails, doctorfis notes, lab tests, health insurances, ultrasounds, social media and medications.

Different platforms such as Hadoop can keep and analyze huge mass of information.

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

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his 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* (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*, Vol. 36. Elsevier, 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] 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.

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

[8] 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.

[9] 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]* 33 (2007), 1–13.

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

[11] 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.

[12] Dina Fine Maron. 2015. Has maternal mortality really doubled in the US. *Scientific American* (2015).

[13] 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.

[14] Yasmin H Neggers. 2016. Trends in maternal mortality in the United States. *Reproductive Toxicology* 64 (2016), 72–76.

[15] World Health Organization, UNICEF, et al. 2012. Trends in maternal mortality: 1990 to 2010: WHO, UNICEF, UNFPA and The World Bank estimates. (2012).

[16] 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.

List of Figures

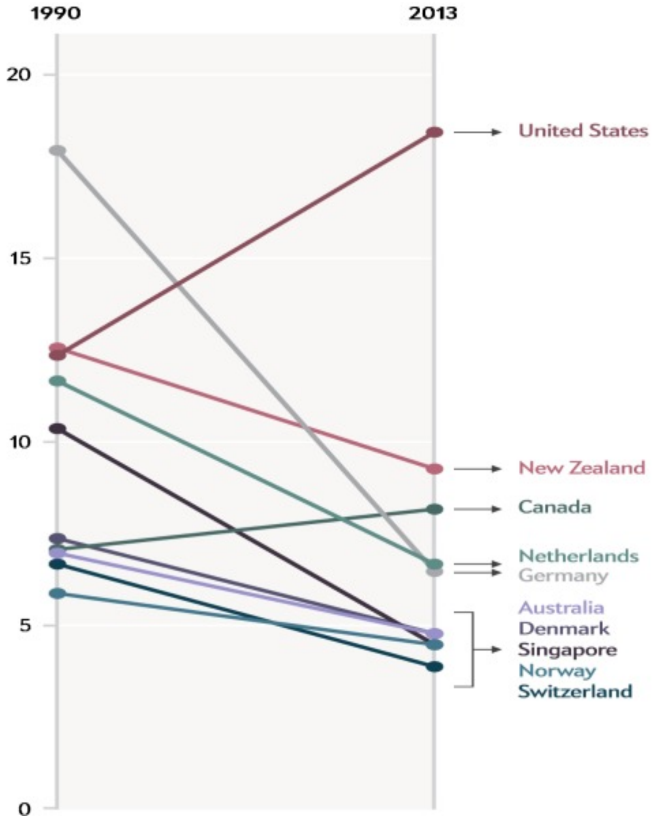


Figure 1: A comparison of maternal mortality ratio in the United States with those of some developed countries between 1990 and 2003.

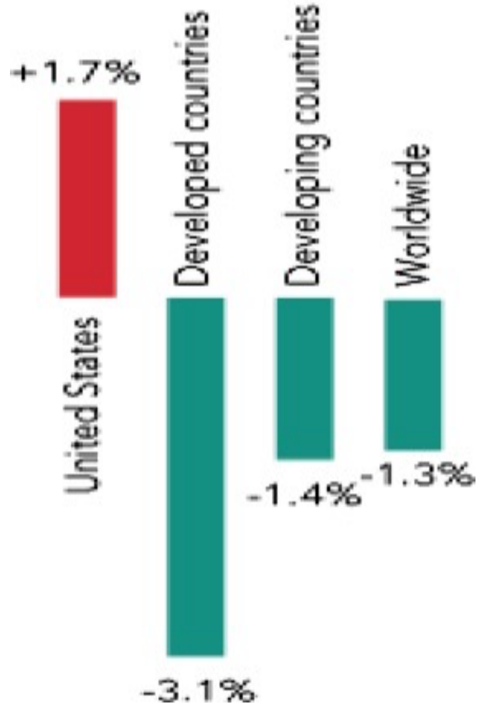


Figure 2: Percentage change in Maternal Mortality Rate between 1990 and 2013 in the United States, worldwide, developed and developing countries.

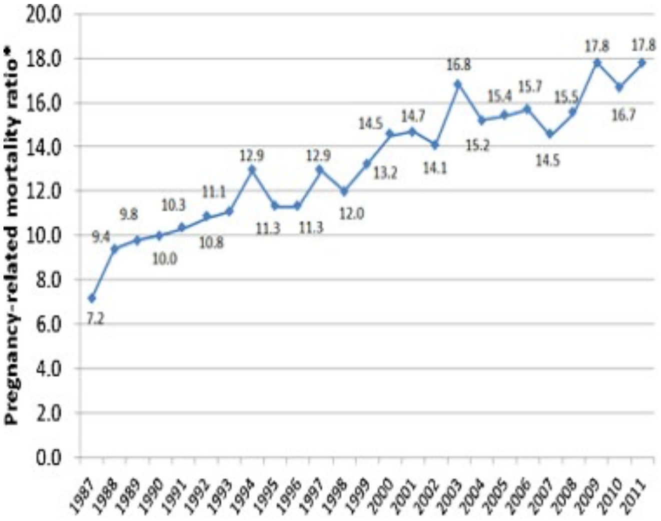


Figure 3: Changes in pregnancy related mortality ratio in United States from 1987-2011.

Impact of Big Data on the Privacy of Mental Health Patients

J. Robert Langlois
Indiana University Bloomington, School of Informatics and Computing
langloir@umail.iu.edu

ABSTRACT

Society has experienced a lot of benefits with the introduction of technology. 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 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.

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 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]. As indicates in [2] ”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.” 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 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. 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." 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].

As found in [1], privacy principles should be introduced during the process of data architecture; privacy should be incorporated into the design and operational procedures. 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 health-care 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 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 CONCLUSION

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 access 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 healthcare 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.

A HEADINGS IN APPENDICES

the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 The Body of the Paper

A.2.1 Type Changes and Special Characters.

A.2.2 Math Equations.

Inline (In-text) Equations.

Display Equations.

A.2.3 Citations.

A.2.4 Tables.

A.2.5 Figures.

A.2.6 Theorem-like Constructs.

A Caveat for the TeX Expert.

A.3 Conclusions

A.4 References

`.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

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).

[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).

[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

This paper will help us to understand about Clinical Trials and how Big data is impacting Clinical Trials. Oncology (Trials) is undergoing a data-driven metamorphosis. Armed with new and ever more efficient molecular and information technologies, we have entered an era where big data is helping us spearhead the fight against various deceases. This technology driven data explosion, often referred to as ”big data”. [3]

KEYWORDS

I523, HID 326, Big data, Clinical, Trials, Health care, Data integration, Analytics

1 INTRODUCTION

A primary objective of clinical trials is gaining knowledge from studying a subset of patients which can then be applied to a much wider group of patients to improve care. In routine practice, patient care is delivered within a rich background of intrinsic and endemic confounding factors and biases associated with practices and patients. [2]

The data collected around the world from various patients, deceases form big data (collection of large data sets). Big data is currently being used on a limited basis in the clinical trials arena, but experts believe its widespread use is coming in the near future. Some hail the great promise it holds in furthering drug discovery. Others are skeptical that it will bring much value and say that enthusiasm should be tempered. [5]

According to IBM, 2.3 trillion gigabytes of data are created every dayfi!?so much that 90% of the data in the world today has been created in the last two years alone. Digital Universe estimates that by 2020, there will be 5,200 gigabytes of data for every man, woman and child on Earth.

It is predicted that the market for Big Data technology and services will reach $16.9 billion in 2015, 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. [6]

2 BIG DATA & CLINICAL RESEARCH

Discovering clinical trials hidden patterns and associations within the heterogeneous data, uncovering new bio markers and drug targets. Allowing the development of predictive disease progression models. Analyzing Real World Data (RWD) as a complementary instrument to clinical trials, for the rapid development of new personalized medicines. The development of advanced statistical methods for learning causal relations from large scale observational data is a crucial element for this analysis. [4]

2.1 Data Integration

Having access to consistent, reliable, and well linked is one of the biggest challenges facing pharmaceutical clinical trials. The ability to manage and integrate data generated at all phases of the value chain, from discovery to real-world use after regulatory approval, is a fundamental requirement to allow companies to derive maximum benefit from the technology trends. Data are the foundation upon which the value-adding analytics are built. Effective end-to-end data integration establishes an authoritative source for all pieces of information and accurately links disparate data regardless of the source be it internal or external, proprietary or publicly available. Data integration also enables comprehensive searches for subsets of data based on the linkages established rather than on the information itself. ”Smart” algorithms linking laboratory and clinical data, for example, could create automatic reports that identify related applications or compounds and raise red flags concerning safety or efficacy. [2]

Implementing end-to-end data integration requires a number of capabilities, including trusted sources of data and documents, the ability to establish cross-linkages between elements, robust quality assurance, workflow management, and role-based access to ensure that specific data elements are visible only to those who are authorized to see it. Pharmaceutical companies generally avoid overhauling their entire data-integration system at once because of the logistical challenges and costs involved, although at least one global pharmaceutical enterprise has employed a ”big bang” approach to remaking 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.

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. [4]

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. [4]

2.3 Data-driven metamorphosis

Data collected in clinical trials undergoing a data-driven metamorphosis. Armed with new and ever more efficient molecular and information technologies, we have entered an era where data is helping us spearhead the fight against cancer. This technology driven data explosion, often referred to as "big data", is not only expediting biomedical discovery, but it is also rapidly transforming the practice of oncology into an information science. This evolution is critical, as results to-date have revealed the immense complexity and genetic heterogeneity of patients and their tumors, a sobering reminder of the challenge facing every patient and their oncologist . This can only be addressed through development of clinico-molecular data analytics that provide a deeper understanding of the mechanisms controlling the biological and clinical response to available therapeutic options. Beyond the exciting implications for improved patient care, such advancements in predictive and evidence-based analytics stand to profoundly affect the processes of cancer drug discovery and associated 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. [4]

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. [4]

3 CHALLENGES

Big pharma companies typically keep their cards close to the vest because it costs so much to develop a drug throughout its lifetime. From discovery to prescription pad, a typical medication can take twelve years and $4 billion to shepherd through its lifecycle, a significant investment that would be hard to recoup if everyone had the secret to the newest blockbuster pill, especially since only ten percent of drugs ever make it 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 patientfis 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. [4]

4 CONCLUSION

The recent surge in big data initiatives in health care is expected to have a positive impact on clinical trials. Increased standardization of common data elements and nomenclature should assist in streamlined trial design and exchange of data. Standardize between trials and will allow easier multi-study analysis. Standardization and quality improvement efforts go hand in hand with a maturing big data infrastructure providing collateral benefits to data curation for trials. [3]

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. (July 2014). 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] 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

[5] 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

[6] 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, Medicare, Medicaid, FWA, 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 "involves 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".[7] 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.[6] In combination these cost the United States healthcare system 80 billion dollars[4] 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 health-care. In order for as many people 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. "If Provider A is involved in improper billing, it is not uncommon for other providers with which they associate to also be engaged in bad behavior. Thus, many payers will work to analyze connected providers. Information on corporate ownership, billing and management companies, social media interactions of physicians and staff can reveal whether other physicians, pharmacies, radiology centers, home infusion agencies, etc. are engaged in a broader pattern of referral and collusion."[8]

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[5]. In addtion 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 5 shows sources of information and what level of FWA they are generally related to. The highest level combines sets of data. "Level 7 combines all previous data views and concerns all fraud that is part of criminal networks which involve many different beneficiaries and/or providers. This much larger data view, spanning billions of claims in the case of Medicaid, is the most rich, delivering the ability to perform complex network analysis that could detect intricate conspiracies. However, performance of analysis here will be much lower than in previous levels."[9]

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. "Much of the FWA that plague health care payers is the result of organized, sophisticated and collusive activities among providers and between providers and patients. Social network analysis can help identify relationships, links and hidden patterns of information sharing and interactions within potentially fraudulent clusters, including:

- Patient relationships with known perpetrators of health care fraud;
- Links between recipients, businesses, assets and relatives and associates;
- Links between licensed and non-licensed and sanctioned providers; and
- Inappropriate relationships between patients, providers, employees, suppliers and partners"[3]

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. "There are several supervised fraud detection methods such as: Bayesian Networks, Neural Networks (NNs), Decision Trees, and Fuzzy Logic. NNs and decision trees are the most popular fraud detection methods because of their high tolerance of noisy data and huge data set handling."[1] There are also unsupervised methods in which data is fed into the system without preexisting notions of what to look for[1]. 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 unsupervisord methods only identify outliers, applying unsupervised methods to healthcare data will require that the outliers will then have to be verified as FWA or acceptable patterns. "Patrick McIntyre, SVP of Health Care Analytics at Anthem, one of the country's biggest payers, credits machine learning and big data with their ability to "identify potentially fraudulent or wasteful claims on a daily basis." The algorithms are run at the same time as claims are batch processed, so questionable claims are immediately identified, flagged and sent to the clinical coding experts for review."[2] 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 patters are discovered and verified.

2.2 Future uses of Big Data Analytics

Currently there is still a certain amount of honor built into healthcare. "The system's inherent structure of trust enables both simple billings errors and illicit actors to hide in the shadows of the murky deep as overpayments quietly siphon money away from legitimate care."[8] 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 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. "By combining identity and entity resolution, rules-based claim and clinical review, complex linking analysis and predictive analytics into a seamless workflow, we will come closer to migrating an integrated pre-pay fraud solution to a real risk control environment with the potential to eliminate billions of dollars in improper payments due to FWA. This is not just a health care imperative, but a national economic imperative that must be addressed immediately. The analytics exist. It is time for those analytics to be implemented and the hard choices that enable that implementation to be made to insure that we remain at the forefront of quality care for all Americans."[3]

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 like 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] 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.

[2] 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/

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

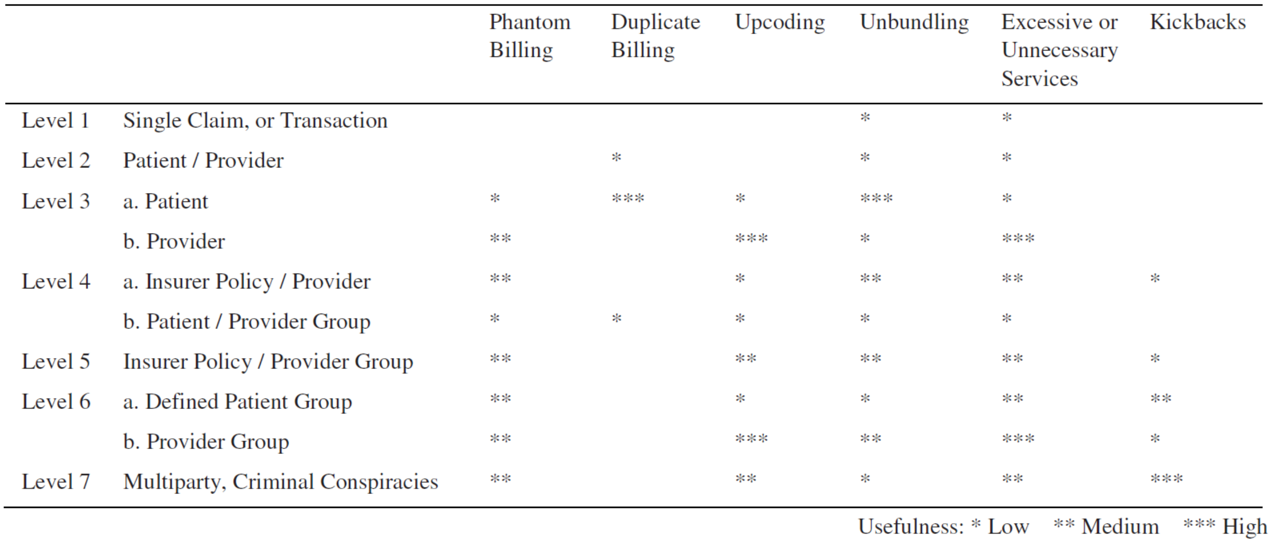
[8] Rodger Smith. 2016. Using Big Data in the Hunt for Healthcare Fraud, Waste, and Abuse Payers must leverage all the big data analytics tools at their disposal to hunt down healthcare fraud, waste, and abuse. Online. (04 2016). https://revcycleintelligence.com/news/using-big-data-in-the-hunt-for-healthcare-fraud-waste-and-abuse

[9] 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.

[Table 1 about here.]

List of Tables

Table 1: Types of Fraud and their related Sources[9]



		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

Big Data Applications in Improving Patient Care

Janaki Mudvari Khatiwada
University of Indiana
Bloomington, Indiana 47408
jmudvari@iu.edu

ABSTRACT

This paper will broadly identify the applications of big data for improving patient care. It will explore 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. It will be an overview of ongoing practices on how big volume of health care data be an important resource for better in-patient and out-patient care.

KEYWORDS

Big Data, Health Care, Patient care Electronic health records

1 INTRODUCTION

Health care is one of the service sector where service providers claim to have provided consumers with the best experience possible, whereas consumers are always browsing for the best care facilities that they could possibly get which might save them time and money and have a quality of life. Health service providers collect high volume of information from the consumers every time they visit the facilities. The volume of health related informations generated in a high velocity of time 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/she spends on it on a weekly or daily basis, general daily meal the person intakes, records on wearable health devices and so on. In today's world big data has become very impactful in policy making, solving problems and making prediction on whole range of areas and health care has become one of the most important sector 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 require, 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 [?]. So to speak fields big data applications in health care for the purpose of improving patient care is wide; disease prevention and management, health education, research and development, prognosis information sharing, public and individual health management, medical optimization. Consumers on the other hand use service provider's websites or web-pages to have an insight of the facilities and the physicians. We look for the ratings and reviews in general based upon which we choose the facility and physician based on other people's experiences for best possible outcome.

2 APPLICATIONS

Health data are stored as electronic medical records(EMR),electronic health records(EMR) 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[?]. One of the trending example of application of big data in tackling opioid crisis in US. Data scientists at Blue Cross Blue Shield have started working with big data experts at Fuzzy Logix to tackle the problem. Using years of insurance and pharmacy data, Fuzzy Logix analysts have been able to identify 742 risk factors that predict with a high degree of accuracy whether someone is at risk for abusing opioids[?]. 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[?]. Johns Hopkins University (Baltimore, MD, USA) developed a disease prediction system using the social media service Twitter[?] The Seton Healthcare Family (Austin, TX, USA) and IBM Joint Development Program have analyzed and tracked medical information, and have predicted outcomes of two million patients per year[?]. Prediction models are especially useful in explaining 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[?].

2.2 Prevention

The mantra, ”Prevention is always better than cure” is what everybody wants to implement. Till now physicians have been studying the general pattern of people's lifestyle and make a recommendation 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 theirweight 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 the trend and prepare them for necessary remedies. Often by partnerships between medical and data professionals, with the potential to peer into the future and identify problems before they happen[**?**]. 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 tailored healthcare package[**?**].

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 researchers by using big data help in early diagnosis.Predictive modeling over data derived from EHRs is being used for early diagnosis and is reducing mortality rates from problems such as congestive heart failure and sepsis[**?**].

2.4 Disease Management

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[**?**].Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient[**?**]

2.5 Research and Development

Through research of big data from past help physicians identity 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.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[**?**]. Big data helps Pharmaceuticals reduce cost of research and therefore lowers drugs cost which benefits patients.

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 challenge. In February, 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. Fortunately they only took identity information such as names and addresses, and details on illnesses and treatments were not exposed[**?**]. 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.

4 CONCLUSIONS

ACKNOWLEDGMENTS

I would like to acknowledge prof.Gregory Von Laszweski, TAs who helped me throughout my writing.

REFERENCES

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

My abstract will go here

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

My introduction will go here [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

SIG Proceedings Paper in LaTeX Format

Jeramy Townsley
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

Sociological Applications of Big Data

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 disciplinefi!?describing and theorizing all of societyfi!?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.

This paper will look at four issues. 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? Fourth, what tools do sociologists have for studying big data? In the process of looking at these questions, applications of big data by sociologists will be described.

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 fibig datafi plus either fisociologyfi or fisocial science.fi There were only 559 in 2005, a mere 7.4% of the total fibig datafi 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 artefact 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.

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 fibig data,fi according to a Google Scholar search (10/5/2017). This is a total search of any timeframe, and as above, not all of these references represent primary research using big data, but simply refer to the term. 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 fibig datafi 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 fi 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.fi 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 was an important resource for this paper.

Big data has famously been described has having velocity, variety and volume [?]. Kitchin (2014) 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). 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 fifitfi based on these differentiating concepts. 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, fireal time flow of data across a whole systemfi that produces a large dataset. The other descriptive concepts are still relevant, and may be pertinent to some big datasets, but not to others.

This is an example on how to refer to Figure 1

[Figure 1 about here.]

ACKNOWLEDGMENTS

temp

REFERENCES

List of Figures

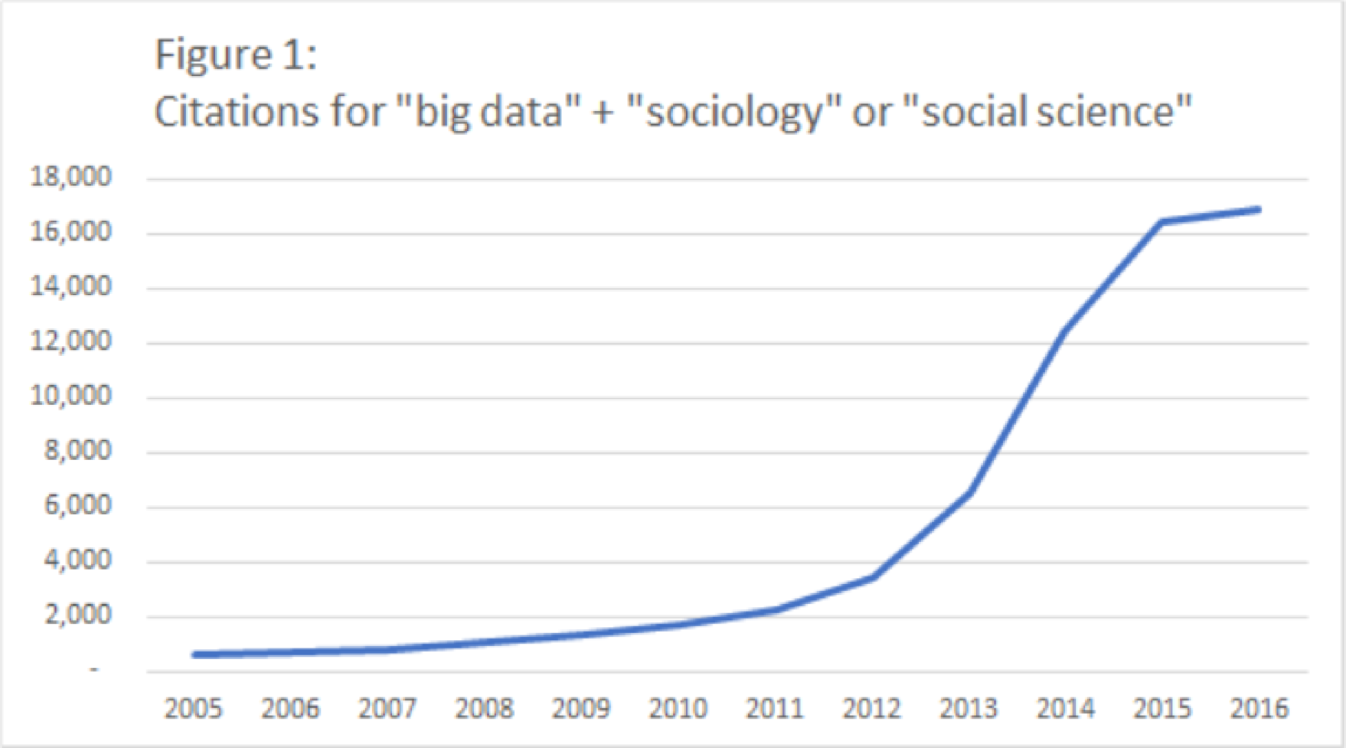


Figure 1: A sample black and white graphic that has been resized with the `includegraphics` command.

Big Data and Deep Learning

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

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

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 Networks has an upwards of Million parameters to train[2]. The 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. I highlight approaches which use cheaper commodity systems integrated in a distributed fashion to handle training such Deep Neural Networks.

KEYWORDS

I523, Distributed Systems, Convolutional 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. They have achieved unparalleled success in the fields of Computer Vision, Natural Language Processing and Artificial Intelligence[3,4]. The current state of the arts Convolutional Neural Networks have billions of parameters to train [5]. Large amounts of data is required to train these parameters. Deep Neural Networks are inherently parallel in nature, with weights and gradient updates shared across layers within the network. Section 2 discusses various ways to introduce parallelism while training Deep Neural Networks. Section 3 discusses methodologies to update Model parameters when the data to train is distributed across multiple machines within the network. Section 4 introduces methodologies to train multiple layers of the same network in parallel in a distributed fashion. Section 5 discusses methodologies and benefits to train multiple parallel networks simultaneously when the required resources are available.

2 PARALLEL AND DISTRIBUTED ARCHITECTURES

2.1 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 projecting 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 various higher level information regarding 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. Recent advanced software architectures such as tensorflow [4] and theano [5] 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.

2.2 Need For Distributed approaches

Standard 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.
- Layer Parallelism - Recent research in Deep Convolutional Networks is focused on the 'wider' paradigm instead of the traditional 'deeper' paradigm [6,7]. Wider Convolutional Networks can be viewed as a stack of smaller networks connected in parallel.These smaller networks can be trained in parallel across multiple cores as these networks do not suffer from resource sharing.
- 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 at different machines.

3 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 ordered 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

3.1 Synchronous Updates

Zinkevich, Weimer, Smola & Li, 2010 [8] introduced a parallel variant of the traditional Stochastic Gradient Descent algorithm. They designed a simple yet efficient algorithm 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 Parallel SGD ($\{c^1, ..., c^m\}, T, n, w_o, k$)

```
for machine ∈ {1....k} in parallel do
    $v_i = SGD(\{c^1, ...., c^m\}, T, n, w_o)$
$v = \frac{1}{k} \sum_{i=1}^{k} v_i$
$Return\ v$
```

3.2 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. Each model requests 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 Downpour SGD (p, d)

```
for machine ∈ {1....k} in parallel do
    $query\ updated\ parameters\ from\ server$
    $v_i = SGD(p, d)$
    $Update\ centralized\ server\ with\ v_i$
    $p = p - \nabla v_i$
```

4 PARALLELISM OF NETWORK LAYERS

Recent research in Deep Convolutional Networks is focused on the 'wider' paradigm instead of the traditional 'deeper' paradigm [6,7]. 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.

5 PARALLELISM OF NETWORKS

Different layers within the Neural Network share their weights and biases. Thus entire copies of The Neural Networks can be trained in parallel with different data inputs fed to the different networks. This significantly reduces the time it takes to train Deep Networks. Access to High Performance Computing resources is required to setup such an environment, this is one of the biggest drawbacks of training several full length networks in parallel.

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 (NIPS'12).* Curran Associates Inc., USA, 1223–1231. http://dl.acm.org/citation.cfm?id=2999134.2999271

[2] 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

Big Data and Artificial Neural Networks

Bharat Mallala
Indiana University
Smith Research Center
2805 E. 10th St, Suite 150
Bloomington, IN 47408, USA
bmallala@iu.edu

ABSTRACT

Big data is often referred as a problem of dealing with large data sets. With the advancements in computational science and the recent evolution of Artificial Intelligence(AI) and Machine Learning, huge volumes of data is being generated every day. Simultaneously the computational resources needed to process and analyze this data is trying to catch up with the rapidly growing data and for the most part have succeeded. In today's world there is a large dependency on Neural networks for dealing with problems in AI and Data analysis. This paper addresses how Big data and its applications can be used to addresses various issues that arise with Artificial Neural Networks(ANN).

KEYWORDS

Artificial Neural Networks, Machine Learning, Artificial Intelligence, Data Analysis, Perceptron.

1 INTRODUCTION

Artificial Neural Networks are often referred as a Multi-layer Neural Network where each node in the network is a Perceptron. It often mics the human brain, i.e. it works in a similar fashion. Advancements in ANN's and its ability to solve complex problem at a relatively faster rate than the traditional approaches have made it the top choice for solving the usually NP-hard AI problems. ”Visual analysis systems will all require a neural network behind them, and that involves a lot of compute power”[**?**] quoted Anderson. This explains the efficiency of Neural networks in solving problems and analysis. ANN's take a series of inputs from the users and map them accordingly to find reasonable patterns in data.

Certainly with these advancements comes huge volumes of data which needs to processed efficiently. This is where Big data comes into picture with its ability to store and process large data sets of any kind for example audio, video, images,text etc in relatively less time. ”s Big Data Analytics is an effective and capable way to, not only work with these data, but understand its meaning, providing inputs for assertive analysis and predictive actions.”[**?**] quotes Victor P Barros in paper.

Artificial Neural Networks usually consists of three primary layers, input layer, output layer, hidden layer. There may be multiple layers of perceptrons within the hidden layer. From the figure 1 we can see the three layers of the ANN. The input layer takes in the input as a set of features and its corresponding weights and the output layer returns a predicted value. All the calculations are done in the hidden layer. The ANN's typically use the feed forward algorithm combined with back propagation for its calculation. The network initially feeds forward to the very end and generates an output from the initial set of features and weights. It then back propagates using Gradient descent and recalculates the wights for each iteration. The algorithm finally stops of the difference in weights form one iteration to the other is not greater than a pre defined threshold. We then test this on the training set and evaluate the performance of the network.

2 CONCLUSIONS

This is my Conlusion

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for all the help he has provided for this paper.

REFERENCES

My First paper

ZhiCheng Zhu
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221

ABSTRACT

This paper edit by zzc

KEYWORDS

info523 big data

1 INTRODUCTION

this is the introduction

2 THE BODY OF THE PAPER

this is the body of the paper

3 CONCLUSIONS

This is the conclusion

ACKNOWLEDGMENTS

this is the acknow of th para

REFERENCES

Big Data Application in Web Search and Text Mining

Wenxuan Han
Indiana University Bloomingtonn
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. In this paper, we first study the data analytic steps in web search, then analyze some of the popular approaches or algorithms (e.g. Hubs, PageRank, etc), and at last, we discuss 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 [3]. 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 [3]. 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 we 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 [4]. The information may be a mix of web pages, images, and other types of files. Since web search is applying on web, 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 [2]) is semi-automatic process of discovering information, meaningful contents, topics, word, relations and patterns from a large amount of text data [3], 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

DIKW is the main part pf we. it the combination of Data, Information, Knowledge and Wisdom. For each element, it has the meaning. Data: Raw web pages or ”Dowuments views as a bag of words” Information: Result of query or ”Dowuments viewed as a collection of insight”

[Figure 1 about here.]

3 TEXT MINING

3.1 lala

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Perrin A. 2015. Social Networking Usage: 2005-2015. (Octobe 2015).
[2] 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).
[3] Mehmet U. and Secren G. 2016. Text Mining Analysis in Turkish Language Using Big Data Tools. *IEEE Computer Society* (2016).
[4] Wikipedia. 2017. Web search engine. (October 2017). https://en.wikipedia.org/wiki/Web_search_engine

List of Figures

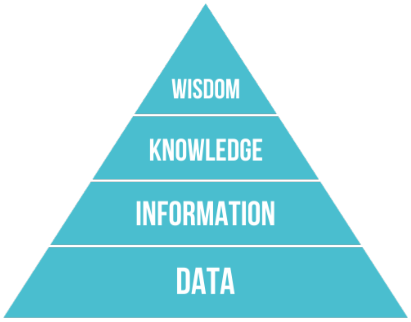


Figure 1: DIKW model.

Using Big Data for Fact Checking

Karthik Vegi
Indiana University
2619 E. 2nd St, Apt 11
Bloomington, IN 47401, USA
kvegi@iu.edu

ABSTRACT

This paper intends to discuss how Big Data can be used to spot fake news, bad data used by politicians, advertisers, and scientists.

KEYWORDS

Big Data, Fact checking

1 INTRODUCTION

Big Data can be used to spot fake news, bad data used by politicians, advertisers, and scientists.

2 CONCLUSIONS

Add a conclusion here

3 REFERENCES

Generated by bibtex from your `.bib` file. Run latex, then bibtex, then latex twice (to resolve references) to create the `.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

ACKNOWLEDGMENTS

I thank all the people who made this possible

REFERENCES

Big Data Applications in 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 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 media and entertainment industry.

KEYWORDS

i523, hid233, Big data, Media, Entertainment industry, Technology, Recommendation

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 rate at which we are generating data is rapidly outpacing our ability to analyze it and the trick here is to turn these massive data streams from a liability into a strength" [2]. Despite that, the technologies used to collect, analyze and interpret data are continuously improving [6].

IDC, the International Data Corporation, believes that "organizations that are best able to make real-time business decisions using big data streams will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure. This will be particularly true in industries experiencing high rates of business change and aggressive consolidation" [7].

But what 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 growing challenge that organizations face as they deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems. Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis [7].

Emerging sources for big data include industries that are preparing to digitize their content. Particularly, "the media and entertainment industry moved to digital recording, production, and delivery in the past five years and is now collecting large amounts of rich content and user viewing behaviors" [7].

2 CHALLENGES IN MEDIA AND ENTERTAINMENT INDUSTRY

"The problem with the massive data collection and distribution system created is: big data is a big mess. Most of the data captured in our daily lives just sits around, cluttering up storage space on devices and slowing down connections" [6].

"Media and entertainment industry has frequently been at the forefront of adopting new technologies. The key business problems that are driving media companies to look at big data capabilities are the need to reduce the costs of operating in an increasingly competitive landscape and at the same time, the need to generate revenue from delivering content and data through diverse platforms and products" [4].

Traditional TV media are facing challenges as its data is scattered. "It has internal data from set-top boxes, network management systems, BOSS systems, etc. as well as external data from online user behaviors. Data integration is the primary challenge in big data applications of traditional TV media" [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. In addition, traditional TV operation rate has decreased from 70 percent in 2009 to 30 percent in 2012" [9].

These are the main challenges media and entertainment industry needs to deal with in order to better utilize big data to make a difference:

- "Making sense of data streams, whether text, image, video, sensors, and so on. Sophisticated products and services can be developed by extracting value from heterogeneous sources" [4].
- "Exploiting big data step changes in the ability to ingest and process raw data, so as to minimize risks in bringing new data-driven offerings to market" [4].
- "Curating quality information out of vast data streams, using algorithmic scalable approaches and blending them with human knowledge through curation platforms" [4].
- "Accelerating business adoption of big data. Consumer awareness is growing and technical improvements continue to reduce the cost of storage and analytics tools among other things. Therefore, it is more important than ever that businesses have confidence that they understand what they want from big data and that the non-technical aspects such as human resources and regulation are in place" [4].

3 APPLICATIONS IN MEDIA AND ENTERTAINMENT INDUSTRY

"Massive quantities of data are already being captured about entertainment. For example, *Supernatural*, an American horror series, created by Eric Kripke in 2005. Now in its seventh season, it has generated roughly 112 hours of footage. We have a lot of pixels and we also have every action of every character; every line of dialogue; a history of when, where, and how often everyone dies. Because all of that information is data, what we actually have, in and around those 112 hours of pixels, is a map to the world of *Supernatural*, and the characters inside it. Today, all of that footage and all of that information is locked away in old style data collections: fixed and unwieldy. But if we can store all that information in a system, modeled more on biology than books, and apply our significant and increasing processing power to analyze and respond to the world, rather than just move it around mechanically, then we have the possibility of generating and interacting with the world and the characters of *Supernatural*" [6].

"Hollywood also uses big data big time" [5]. "IBM worked with a media company and ran its predictive models on social media buzz for the movie Ram Leela. According to the reports, IBM predicted a 73 percent success for the movie based on right selection of cities. Such rich analysis of social media data was conducted for Barfi and Ek Tha Tiger. All these movies had a runaway success at the box office. Shah Rukh Khan's Chennai Express, one of the biggest box office grosses in 2013, used big data and analytic solutions to drive social media and digital marketing campaigns. IT services company *Persistent Systems* helped Chennai Express team with the right strategic inputs. Chennai Express related tweets generated over 1 billion cumulative impressions and the total number of tweets across all hash tags was more than 750 thousand over the 90-day campaign period. Singapore based big data analytic firm Crayon has worked with leading Hindi film industry producers to understand the kind of music to release in order to create the right buzz for movies. In addition, Lady Gaga and her team browse through listening preferences and sequences to optimize the play list for maximum impact at live events" [3].

"Sports is another area where big data is making big impacts. Germany, FIFA 2014 champion, has been using SAP's Match Insights software to analyze team performance which made a big difference for the team. It analyzes data such as player positions, touch maps, passing abilities, ball retention and even metrics such as aggressive play. In addition, Kolkata Knight Riders, an Indian Premier League team, used Match Insights to determine the consistency of its players which helped in auction as well as in ongoing training" [3].

"By using big data to understand why the customers subscribe and unsubscribe, entertainment organizations could develop the best product and promotional strategies to attract and retain clients. Unstructured sources best handled by big data apps like email, call detail records and social media sentiment reveal factors that are often overlooked for driving customer interest. Big data makes possible the understanding of consumption of digital media and entertainment and behavior that could be used together with traditional data demographic for personalized advertising in the right context at the right time, in the right place" [5].

"Recommendation engines are very powerful personalization tools because it's a great way to show people items they will like. A lot of Amazon's fantastic revenue growth has been built on successfully integrating recommendations across the buying experience from product discovery to checkout" [1].

"Amazon is investing a large amount of talent and resources on getting better artificial intelligence, specifically deep learning technology to make recommendation engines which learn and scale even more efficiently. In May 2016, Amazon opened up its sophisticated artificial intelligence technology as a cloud platform. The company unveiled DSSTNE, an open source artificial intelligence framework that Amazon developed to power its own product recommendation system" [1].

"Amazon says it releases pilots at Amazon Studios periodically for customers to watch and review. Their feedback is taken into account when executives decide which pilots will become a full series. One product of that system is the comedy series *Transparent*, based on a Los Angeles family whose patriarch is transgender. Its debut in 2014 coincided with greater social awareness about transgender issues and 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. "No one understands the idea of content discovery better than Netflix, because the on-demand streaming video is probably the world's biggest market for digital consumption of content. Netflix has worked hard to ensure its recommendation algorithms can highlight as much of its content library as possible. In December 2015, Netflix revamped the technology behind its content recommendation engine, deciding to do away with region based preferences in light of their ongoing global expansion" [1].

"Netflix, which distributes shows such as *House of Cards* and *Orange Is the New Black*, pioneered the use of mathematical equations to promote titles that a subscriber might enjoy. That is based on variables such as previously downloaded content, the subscriber's location and the show's broader popularity" [8].

"A typical Netflix user may lose interest unless something interesting is found within 60 seconds, two employees of the Los Gatos, California-based company wrote in a paper published in a scholarly journal last year. Netflix's system for coming up with personalized viewing recommendations helps save more than 1 billion dollar a year by reducing the number of subscription cancellations" [8].

4 CONCLUSION

The rapid growth of big data has given 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.

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

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. This material is 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."The goal of a recommender system is to 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 is 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 tweaked to help people find videos to watch or news to read, been challenged by other algorithms and other techniques, and been adapted to improve diversity and discovery, recency, time-sensitive or sequential items, and 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 recommendations aim to "leverage all the available user data provided in various systems and domains, in order 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 used in this paper.

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 now face serious challenges in serving the evolving needs of researchers, which are increasingly focused on solutions for preserving and processing large amounts of data. To remain relevant in the current information landscape, libraries and archives must implement new strategies for converting legacy data to formats that can add value to the research lifecycle.

KEYWORDS

Libraries, Archives, Data Management, Data Integration, ETL

1 INTRODUCTION

Examples of big data analytics in research libraries and archives are still scarce. In the library domain, the leading data hub is the Online Computer Library Center (OCLC)[1].

2 CONCLUSION

Conclusions and abstracts must not have any citations in the section.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 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

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Rono-Hills
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
30 Shuangqing Rd
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

2 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

A HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e., the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 The Body of the Paper

A.2.1 Type Changes and Special Characters.

A.2.2 Math Equations.

Inline (In-text) Equations.

Display Equations.

A.2.3 Citations.

A.2.4 Tables.

A.2.5 Figures.

A.2.6 Theorem-like Constructs.

A Caveat for the TeX Expert.

A.3 Conclusions

A.4 References

Generated by bibtex from your `.bib` file. Run latex, then bibtex, then latex twice (to resolve references) to create the `.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

B MORE HELP FOR THE HARDY

Of course, reading the source code is always useful. The file `acmart.pdf` contains both the user guide and the commented code.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Rono-Hills
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
30 Shuangqing Rd
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

Big Data Analytics in Biometric Identity Management

Robert W. Gasiewicz
Indiana University
711 N. Park Avenue
Bloomington, IN 47408
rgasiewi@iu.edu

ABSTRACT

This paper is intended to be a primer for understanding how 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

REFERENCES

[1] Simon A. Cole. 2002. *Suspect Identities: A History of Fingerprinting and Criminal Identification*. Academic Trade. (book).

[2] Information Sharing Environment. [n. d.]. Executive Order 13388. ([n. d.]). 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. 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. (book).

[5] U.S. Marshals Service Website. [n. d.]. Fingerprint History. ([n. d.]). 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

Having an intelligent ear-and-eye monitoring system at the home to constantly observe the surroundings both inside and outside can protect the house and personnel much more safer way. By extending this capability to the neighborhood and city through collaboration would create safe cities across the world.

Anti-social activities became the most significant threat to national security because of their potential to bring massive damage to our homes, public infrastructure, economy, and people. The existing systems and methods havenfit reached the level of sophistication to be able to consolidate the relevant data from variety of sources and demographics. Video surveillance of residential, commercial, military, and other restricted locations have been in practice since many years using various available technologies. Depending on the level of security, the data has been processed by data mining and/or big data analytics to take decisions by various personal, agencies and governments. However, the limitations of data collection, data mining and adoption of intelligence led to ineffective systems which are not predictive as they should be. With the advent of technologies, it possible to integrate the video, audio and social media data of targeted regions (homes, public places and extended areas) for security analysis. Such systems can use advanced statistical methods, classification algorithms and machine learning algorithms to predict and prevent the threats based on the severity probability.

KEYWORDS

i523, HID333, HID337, Artificial Intelligence, Natural Language Processing (NLP), Machine Learning, Micro Drone

1 INTRODUCTION

As per the book ”Intelligence and Security Informatics”[1], it is widely believed that information technology will play an indispensable role in making the world safer by supporting intelligence and knowledge discovery through collecting, processing, analyzing, and utilizing terrorism- and crime-related data. Social network analysis(SNA)has been widely explored to support intelligence and law enforcement agencies in investigating the terrorist and criminal social networks. It is valuable in identifying terrorists, suspect subgroups, and their communication patterns.

However, in the present world, the systems are disparately processing the data and the decisions/conclusions are being made without considering from multiple dimensions. The large corporations, nations, and intelligence agencies are using their individual systems but not taking integrated approach to solve the problems in their entirety. This could be due to political and economic interests of individuals, corporations and nations.

Analyzing the 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 research.

It is quite feasible to collect the 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 governments and law enforcement agencies. By accessing the lowest level of data of individualsfi video, voice, social media and other business transactional data would allow to characterize, analyze and assess the people behaviors and motives which can be maintained and processed as needed by Big Data systems. These systems are very complex in nature due to the variety, volume and velocity of the data, where the Big Data 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, it has been described some of the technologies/sub-systems that can be used to achieve the objectives of proposed conceptual model.

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. The remainder of this discussion is organized as follows: Section 2 provides a set of sub-systems and use of them in the proposed framework. Section 3 gives an overview community/city security using conceptual model. Finally, section 4 concludes this proposal.

1.1 Data Collection

As discussed in the above sections, the data can be collected from individual home level using the following means: video and voice enabled micro-drone, re-chargeable dock with edge processing server, WLAN (both infrastructure and Adhoc mode) with ability to transfer data to cloud server. This miniature sub-system hardware will have capability to harvest the large amounts of data from all major social media accounts of individual house mates such as Twitter, Facebook, YouTube, Instagram, WhatsApp, and Mobile

Phone Calls and Texts and transfer to Big Data infrastructure. The Big Data infrastructure would organize the data through multiple data layers such as collection hub, staging hub and data lake.

1.2 Data Privacy

In the conceptual model that we have discussed, multiple subsystems collect the individual home level information and uses anonymization models to preserve the privacy of individuals. The objective is hiding the sensitive personal information such as personal identities but publishing the rest of the data, an anonymized version of relational data. The data that will be sent out to be used for next level (community/region) fed to privacy preservation algorithms such as k-anonymity protection models which are being used in real-world systems known as Datafly, -Argus and k-Similar. The k-anonymity methods ensure that at least k records with respect to every set of quasi-identifier attributes are indistinguishable. There are other alternative methods such as l-diversity and m-invariance can be applied as well to apply different constraints on anonymity. For social network integration in to 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” [3].

1.3 Video Analytics

The high quality video image frames will be processed to analyze the 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 object detection, object tracking, face recognition, image classification and scene labeling would enable to establish a comprehensive situational awareness in the home security context. For example, facial expressions manifest not only emotions but also allied actions, behavioral patterns and give a lot of useful data when it comes to helping law enforcement and forensics agencies. Video Analytics can be achieved based on data curation, sentiment analysis, and other advanced solutions. Expressions like ”happy”, ”sad”, ”angry”, ”scared”, ”surprised” or ”neutral” form the basis of video analytics.

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 management, vehicle management, behavior monitoring. For example, in the public events deep learning enabled systems can perform crowd detection, queue management, people counting, people scattering, people tracking; in the vehicle management, systems can perform vehicle classification, traffic monitoring, license plate recognition, road data gathering. Also, behavior monitoring can be achieved through motion detection, vandalism detection, face detection, privacy masking, and suspicious activity detection. 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.

1.4 Voice Analytics

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. Deep learning is becoming a mainstream technology for speech recognition and has successfully replaced Gaussian mixtures for speech recognition and feature coding at an increasingly larger scale. Googlefis Speech Recognition API built using deep learning neural network algorithms is one of the voice analytics software available in the market, which can be used in the proposed conceptual model.

1.5 Social Media Analytics

Security systems can be made more innovative and intelligent by integrating with social data. In the current world, social media like facebook,twitter,Instagram,watsapp etc are generating massive amounts of Big Data.If this information is used intelligently, it can predict behavior patterns and help to know about others.Typically, social analytics are used by marketing departments to predict the behavior and experience that consumers share in social networks. The same data can be used to monitor criminal activities.

2 HOME SECURITY

This section describes a proposed scalable security system building block concept, which can be extended to 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 housemates (e.g., E-mail, Facebook, Twitter, WhatsApp, and other web/mobile applications) gets integrated in to HDFS server.

This will establish a known context of complete information of individuals residing in-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 super knowledge about all the elements. This will require Big Data infrastructure along with 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 scalable system.

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 [4] have reviewed the recent progress of deep learning in object detection, object tracking, face recognition, image classification and scene labeling. The 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.

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 face-voice association, the 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 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 emotions based on the expressions, conversations and actions to predict the immediate 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 (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 reasonably.

The technologies available to realize such a micro-drone consists of: autonomous octo/quadcopters, high resolution built-in 360 degree video cameras, high speed network link, high speed GPUs, environment sensors, software with machine learning algorithms for various capabilities discussed above.

2.2 Learning Algorithms

The two critical machine learning algorithms needed to realize the proposed concept are for the face and voice recognition. Deep learning models are potential candidates for these two tasks. Deep learning architectures have different variants such as Deep Belief Networks (DBN) [4], Convolutional Neural Networks (CNN) [5], Deep Boltzmann Machines (DBM) [6] and Stacked Denoising Auto-Encoders (SDAE) [7], etc. The most attractive model is Convolutional Neural Networks which have achieved very promising results in both computer vision and speech recognition.

2.3 Predictive Analysis

The ability to estimate the occurrence of future events using expertise, observation and intuition is critical to the human decision-making process. From a biophysical perspective, there is strong evidence that the neocortex provides a basic framework for memory and prediction in which human intelligence emerges as a process of pattern storage, recognition and projection rooted in our experience of the world and driven by perception and creativity. There is increasing consensus among cognitive psychologists that human decision making can be seen as a situation-action matching process which is context-bound and driven by experiential knowledge and intuition.

2.4 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 collection of security drones covering a specific region within a city which will ensure that 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 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 CITY/EXTENDED REGIONAL SECURITY

The proposed conceptual model defines 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 the 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'. All the traditional networking principles can be applied to drone network as well with the artificial intelligence enabled capabilities. The information that will be routed from home-to-community-to-city-to-next level follows privacy preservation policies and law enforcement/intelligence agency regulations.

4 CONCLUSIONS

In this discussion, based on the review of existing literature, systems/products and technologies it has been perceived that security informatics systems are disparately implemented and consolidation of data and analysis at various layers hasnfit been done efficiently. Considering that Big Data technologies are robust enough to collect the large volumes of data from variety of sources, a conceptual model is proposed to discuss the feasibility of integrated video, voice, and social media data of individuals, communities and cities to be collected and analyzed to apply the 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

REFERENCES

Big Data Analytics in Sports - Track and Field

Mathew Millard
Indiana University Bloomington
938 N Walnut St. Apt. G
Bloomington, Indiana 47404
mdmillar@indiana.edu

ABSTRACT

This paper covers the impact that Big Data has and could have on the sport of track and field.

KEYWORDS

i523

1 INTRODUCTION

This is my introduction

2 THE BODY OF THE PAPER

This is the body of my paper

3 CONCLUSIONS

This is my conclusion

ACKNOWLEDGMENTS

Acknowledgments

REFERENCES

Big Data Analytics in Sports - Soccer

Rahul Velayutham
Indiana University Bloomington
2661 E 7th Street Apt H
Bloomington, Indiana 47408
rahul.vela@gmail.com.com

ABSTRACT

The aim of this paper is to provide an understanding as to how big data is playing a huge role in Football clubs helping them scout players.

KEYWORDS

Big Data, Soccer , Scouting

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

REFERENCES

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. Footballfis 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

Big Data Analytics using Spark

Nisha Chandwani
Indiana University Bloomington
107 S Indiana Ave
Bloomington, Indiana 47405
nchandwa@iu.edu

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

2 THE BODY OF THE PAPER

Typically, the body of a paper is organized into a hierarchical structure, with numbered or unnumbered headings for sections, subsections, sub-subsections, and even smaller sections. The command `\section` that precedes this paragraph is part of such a hierarchy. LaTeX handles the numbering and placement of these headings for you, when you use the appropriate heading commands around the titles of the headings. If you want a sub-subsection or smaller part to be unnumbered in your output, simply append an asterisk to the command name. Examples of both numbered and unnumbered headings will appear throughout the balance of this sample document.

Because the entire article is contained in the **document** environment, you can indicate the start of a new paragraph with a blank line in your input file; that is why this sentence forms a separate paragraph.

2.1 Type Changes and *Special* Characters

We have already seen several typeface changes in this sample. You can indicate italicized words or phrases in your text with the command `\textit`; emboldening with the command `\textbf` and typewriter-style (for instance, for computer code) with `\texttt`. But remember, you do not have to indicate typestyle changes when such changes are part of the *structural* elements of your article; for instance, the heading of this subsection will be in a sans serif[1] typeface, but that is handled by the document class file. Take care with the use of the curly braces in typeface changes; they mark the beginning and end of the text that is to be in the different typeface.

You can use whatever symbols, accented characters, or non-English characters you need anywhere in your document; you can find a complete list of what is available in the *LaTeX User's Guide* [26].

2.2 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

2.2.1 Inline (In-text) Equations. A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [26]; this section will simply show a few examples of in-text equations in context. Notice how this equation:
$\lim_{n\rightarrow \infty}x=0$,
set here in in-line math style, looks slightly different when set in display style. (See next section).

2.2.2 Display Equations. A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\rightarrow \infty}x=0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty}x_i=\int_{0}^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

2.3 Citations

Citations to articles [6–8, 19], conference proceedings [8] or maybe books [26, 34] listed in the Bibliography section of your article will

[1] Another footnote here. Let's make this a rather long one to see how it looks. Footnotes must be avoided.

occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [26]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide* by Lamport [26].

This article shows only the plainest form of the citation command, using `\cite`.

Some examples. A paginated journal article [2], an enumerated journal article [11], a reference to an entire issue [10], a monograph (whole book) [25], a monograph/whole book in a series (see 2a in spec. document) [18], a divisible-book such as an anthology or compilation [13] followed by the same example, however we only output the series if the volume number is given [14] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [37], a chapter in a divisible book in a series [12], a multi-volume work as book [24], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [4], a proceedings article with all possible elements [36], an example of an enumerated proceedings article [16], an informally published work [17], a doctoral dissertation [9], a master's thesis: [5], an online document / world wide web resource [1, 30, 38], a video game (Case 1) [29] and (Case 2) [28] and [27] and (Case 3) a patent [35], work accepted for publication [31], 'YYYYb'-test for prolific author [32] and [33]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [23]. Boris / Barbara Beeton: multi-volume works as books [21] and [20].

A couple of citations with DOIs: [22, 23].

Online citations: [38–40].

We use jabref to manage all citations. A paper without managing a bib file will be returned without review. in the bibtex file all urls are added to rfernces with the *url* filed. They are not to be included in the *howpublished* or *note* field.

2.4 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

[Table 1 about here.]

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

[Table 2 about here.]

It is strongly recommended to use the package booktabs [15] and follow its main principles of typography with respect to tables:

(1) Never, ever use vertical rules.
(2) Never use double rules.

It is also a good idea not to overuse horizontal rules.

2.5 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of `.eps` files to be displayable with LaTeX. If you work with pdfLaTeX, use files in the `.pdf` format. Note that most modern TeX systems will convert `.eps` to `.pdf` for you on the fly. More details on each of these are found in the *Author's Guide*.

[Figure 1 about here.]

[Figure 2 about here.]

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. And don't forget to end the environment with **figure***, not **figure**!

[Figure 3 about here.]

[Figure 4 about here.]

2.6 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. ACM uses two types of these constructs: theorem-like and definition-like.

Here is a theorem:

THEOREM 2.1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)\,dt = G(b) - G(a).$$

Here is a definition:

Definition 2.2. If z is irrational, then by e^z we mean the unique number that has logarithm z:

$$\log e^z = z.$$

The pre-defined theorem-like constructs are **theorem**, **conjecture**, **proposition**, **lemma** and **corollary**. The pre-defined definition-like constructs are **example** and **definition**. You can add your own constructs using the *amsthm* interface [3]. The styles used in the `\theoremstyle` command are **acmplain** and **acmdefinition**.

Another construct is **proof**, for example, Proof. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

3 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

A HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e., the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 The Body of the Paper

A.2.1 Type Changes and Special Characters.

A.2.2 Math Equations.

Inline (In-text) Equations.

Display Equations.

A.2.3 Citations.

A.2.4 Tables.

A.2.5 Figures.

A.2.6 Theorem-like Constructs.

A Caveat for the TeX Expert.

A.3 Conclusions

A.4 References

Generated by bibtex from your `.bib` file. Run latex, then bibtex, then latex twice (to resolve references) to create the `.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

B MORE HELP FOR THE HARDY

Of course, reading the source code is always useful. The file `acmart.pdf` contains both the user guide and the commented code.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

[1] Rafal Ablamowicz and Bertfried Fauser. 2007. CLIFFORD: a Maple 11 Package for Clifford Algebra Computations, version 11. (2007). Retrieved February 28, 2008 from http://math.tntech.edu/rafal/cliff11/index.html

[2] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. https://doi.org/10.1145/1188913.1188915

[3] American Mathematical Society 2015. *Using the amsthm Package.* American Mathematical Society. http://www.ctan.org/pkg/amsthm

[4] Sten Andler. 1979. Predicate Path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages (POPL '79).* ACM Press, New York, NY, 226–236. https://doi.org/10.1145/567752.567774

[5] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle.* Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[6] Mic Bowman, Saumya K. Debray, and Larry L. Peterson. 1993. Reasoning About Naming Systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 795–825. https://doi.org/10.1145/161468.161471

[7] Johannes Braams. 1991. Babel, a Multilingual Style-Option System for Use with LaTeX's Standard Document Styles. *TUGboat* 12, 2 (June 1991), 291–301.

[8] Malcolm Clark. 1991. Post Congress Tristesse. In *TeX90 Conference Proceedings.* TeX Users Group, 84–89.

[9] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry).* Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[10] Jacques Cohen (Ed.). 1996. Special issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[11] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[12] Bruce P. Douglass, David Harel, and Mark B. Trakhtenbrot. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. https://doi.org/10.1007/3-540-65193-4_29

[13] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

[14] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. https://doi.org/10.1007/3-540-09237-4

[15] Simon Fear. 2005. *Publication quality tables in LaTeX.* http://www.ctan.org/pkg/booktabs

[16] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07).* USENIX Association, Berkley, CA, Article 7, 9 pages.

[17] David Harel. 1978. *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER.* MIT Research Lab Technical Report TR-200. Massachusetts Institute of Technology, Cambridge, MA.

[18] David Harel. 1979. *First-Order Dynamic Logic.* Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. https://doi.org/10.1007/3-540-09237-4

[19] Maurice Herlihy. 1993. A Methodology for Implementing Highly Concurrent Data Objects. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 745–770. https://doi.org/10.1145/161468.161469

[20] Lars Hörmander. 1985. *The analysis of linear partial differential operators. III.* Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. viii+525 pages. Pseudodifferential operators.

[21] Lars Hörmander. 1985. *The analysis of linear partial differential operators. IV.* Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. vii+352 pages. Fourier integral operators.

[22] IEEE 2004. IEEE TCSC Executive Committee. In *Proceedings of the IEEE International Conference on Web Services (ICWS '04).* IEEE Computer Society, Washington, DC, USA, 21–22. https://doi.org/10.1109/ICWS.2004.64

[23] Markus Kirschmer and John Voight. 2010. Algorithmic Enumeration of Ideal Classes for Quaternion Orders. *SIAM J. Comput.* 39, 5 (Jan. 2010), 1714–1747. https://doi.org/10.1137/080734467

[24] Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.).* Addison Wesley Longman Publishing Co., Inc.

[25] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

[26] Leslie Lamport. 1986. *LaTeX: A Document Preparation System.* Addison-Wesley, Reading, MA.

[27] Newton Lee. 2005. Interview with Bill Kinder: January 13, 2005. Video. *Comput. Entertain.* 3, 1, Article 4 (Jan.-March 2005). https://doi.org/10.1145/1057270.1057278

[28] Dave Novak. 2003. Solder man. Video. In *ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003).* ACM Press, New York, NY, 4. https://doi.org/99.9999/woot07-S422

[29] Barack Obama. 2008. A more perfect union. Video. (5 March 2008). Retrieved March 21, 2008 from http://video.google.com/videoplay?docid=6528042696351994555

[30] Poker-Edge.Com. 2006. Stats and Analysis. (March 2006). Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php

[31] Bernard Rous. 2008. The Enabling of Digital Libraries. *Digital Libraries* 12, 3, Article 5 (July 2008). To appear.

[32] Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

[33] Mehdi Saeedi, Morteza Saheb Zamani, Mehdi Sedighi, and Zahra Sasanian. 2010. Synthesis of Reversible Circuit Using Cycle-Based Approach. *J. Emerg. Technol. Comput. Syst.* 6, 4 (Dec. 2010).

[34] S.L. Salas and Einar Hille. 1978. *Calculus: One and Several Variable.* John Wiley and Sons, New York.

[35] Joseph Scientist. 2009. The fountain of youth. (Aug. 2009). Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.

[36] Stan W. Smith. 2010. An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and Computers (LAC '10)*, Reginald N. Smythe and Alexander Noble (Eds.), Vol. 3. Paparazzi Press, Milan Italy, 422–431. https://doi.org/99.9999/woot07-S422

[37] Asad Z. Spector. 1990. Achieving application requirements. In *Distributed Systems* (2nd. ed.), Sape Mullender (Ed.). ACM Press, New York, NY, 19–33. https://doi.org/10.1145/90417.90738

[38] Harry Thornburg. 2001. Introduction to Bayesian Statistics. (March 2001). Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

[39] TUG 2017. Institutional members of the TeX Users Group. (2017). Retrieved May 27, 2017 from http://wwtug.org/instmem.html

[40] Boris Veytsman. [n. d.]. acmart—Class for typesetting publications of ACM. ([n. d.]). Retrieved May 27, 2017 from http://www.ctan.org/pkg/acmart

List of Figures

Figure 1: A sample black and white graphic.

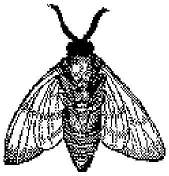


Figure 2: A sample black and white graphic that has been resized with the `includegraphics` command.

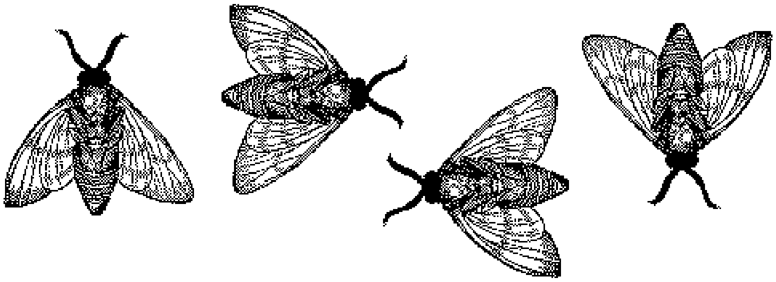


Figure 3: A sample black and white graphic that needs to span two columns of text.

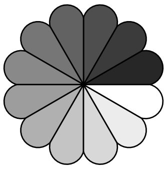


Figure 4: A sample black and white graphic that has been resized with the `includegraphics` command.

List of Tables

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Table 2: Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

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 how they are related to each other and what new technologies are in use in this field.

KEYWORDS

ACM proceedings, LaTeX, text tagging

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.

1.1 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. 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 poses both the greatest challenge and the greatest opportunity as big data could help understand many organizations to understand people better and allocate resources more effectively. Big Data velocity also raises are 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.

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 your search history on any website.

1.2 High Performance Computing

High-performance computing (HPC) is a term used for computers having a capacity of doing 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 don't see much of them anymore. This type of architecture is quite expensive and doesn't really scale.

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).

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 thats 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.

2 THE BODY OF THE PAPER

3 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices;

brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

Big Data Analysis using MapReduce

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

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

REFERENCES

Big Data And Data Visualization

Pravin Deshmukh
Indiana University
300 N. Jordan Avenue
Bloomington, Indiana 47405-1106
praadesh@iu.edu

ABSTRACT

This article provides an overview on importance of data visualization in presenting findings of Big Data solutions.

KEYWORDS

Data Visualization, LaTeX, text tagging

1 INTRODUCTION

Big data is widely used technology to consume huge amount of data. While there are various technologies available to process this data it is very important to have interactive, intuitive, user friendly data visualizations in place so that decision makers, business users will have clear understanding of findings of big data solutions. These visualizations will make help us to make informed decision looking at various trends over the period of time[**?**].

2 CONCLUSION

I like LaTex.

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

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 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 scaleability 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 different LIMS or ELN. This severely restricts data sharing and reuse between groups which leads to many problems including 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 software on-top as well as data created from such software to be accessed and used my a multitude of users[7].

The benefits and challenges of using a PaaS approach to share and regulate R&D data within a large pharmaceutical company that has already implemented numerous laboratory systems will be outlined below.

2 IMPORTANCE OF PLATFORMS

Many organizations struggle with the aim of sharing data and processing tools among researchers. SaaP provides a method of better resource utilization while reducing maintenance costs[6].

3 IMPLEMENTING PLATFORMS

Some of the biggest concerns associated with implementing platforms involve 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 the applications available, the storage solution's design, the cost, the learning curve, or a number of other client based requirements. Data architecture has the overarching purpose to design the data warehouse solution without limitations to growth or analysis tools and 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)[5]. 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, multifactor access control, and biometrics to a limited population of less than 1% of Googlers[3].

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 a 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[4].

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[2].

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 complexity of the 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 they transitions and interconnections still have to be outlined by a designer[8].

A system's user interface has to be laid out in a simple and intuitively manner that allows users to perform the tasks required while exploring new insights provided by the data. There are a number of influences leading to the development of user interfaces such as familiarity where users are use to performing a search in Google or Amazon interfaces and maintain the same high expectation with their working environment. When users track packages with FedEx or UPS they expect the same level of access to sample tracking within their working environment.

4 CONCLUSION: PLATFORMS AND BIG DATA WITHIN THE PHARMACEUTICAL INDUSTRY

As more and more companies realize the value of their data, platforms and the 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 work processes. The 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 then 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 Saber Sheybani and Miao Jiang.

REFERENCES

[1] 2017. Big Data on AWS. Website. (Oct. 2017). https://aws.amazon.com/big-data/

[2] 2017. Big Data Solutions. Website. (Oct. 2017). https://cloud.google.com/products/big-data/

[3] 2017. Google Security Whitepaper. Website. (Oct. 2017). https://cloud.google.com/security/whitepaper#state-of-the-art_data_centers

[4] 2017. Understanding Microsoft big data solutions. Website. (Oct. 2017). https://msdn.microsoft.com/en-us/library/dn749804.aspx

[5] 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* (01 2014), 752–757. https://www.researchgate.net/publication/289573602

[6] 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

[7] 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. 975–984.

[8] Jerome H. Saltzer and M. Frans Kaashoek. 2009. *Principles of Computer System Design: An Introduction.* Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-374957-4.00010-4

Big Data for Edge Computing

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

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

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

Big Data, Edge Computing i523

1 INTRODUCTION

Put here an introduction about your topic. We just need one sample refernce so the paper compiles in LaTeX so we put it here [1].

2 FROM PDF

At the core of Big Data is a challenge. A challenge of exploration fiof the complexities inherently trapped in data, business, and problem-solving systems" (Cao, 2017) [**?**] 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. Mobility network traffic providers generate CDRfis [**?**] (Call Detail Records) every time a device establishes a connection. These CDRs provide details about the connection fi?! cell site locations, length of call and device 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.

As shown in Figure 1 ....

While CDRfis tell us a great deal, there is much that they donfit tell a provider. For example, by the time the millions of records are consolidated to generate files that are more manageable, data details can be lost. Therefore, other data sources are used, like data from the network, which provide precise traffic metrics. Adding to the challenge, companies like Verizon and AT&T are changing to unlimited plans, offering package deals with video services and even offering free traffic based on cell phone apps (HBO for free on your device) fi?! This data set is much different than simply looking at network or CDR traffic. That is, we need to look at the bits and bytes. Additionally, the customer landscape has changed from traditional post-paid customers to those that pre-paid or are sold as wholesale or the IoT customers. With IoT, lots of projections abound and here is one: "roughly 23 billion active IoT devices by the year 2019fi.spending on enterprise IoT products and services will reach $255 billion globally by 2019, up from $46.2 billion this year. " (Schofield, 2015) 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. The impact of apps on the mobility network is significant with no end in site: "fiwhen it comes to reaching consumers in massefithe market has confirmed what wefive known all along fi?! 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." (Shah, 2014) [**?**]

For mobility network providers, what is the Big Data challenge here? Providers already have access to network traffic data, along with data around traffic above the (OSI) network layer 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 readings on - customer behavior! Can it be done? I believe, with the use of data holistically and with data-driven discovery, it can. In order for this to be successful, you have to have a solid understanding of the data itself and 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 fi?! customer behavior. At this point we have gained wisdom and data-driven discovery that can be applied to the network itself. Impacting the bottom line.

[Figure 1 about here.]

3 CONCLUSION

Put here an conclusion. Conlcusions and abstracts must not have any citations in the section.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

List of Figures

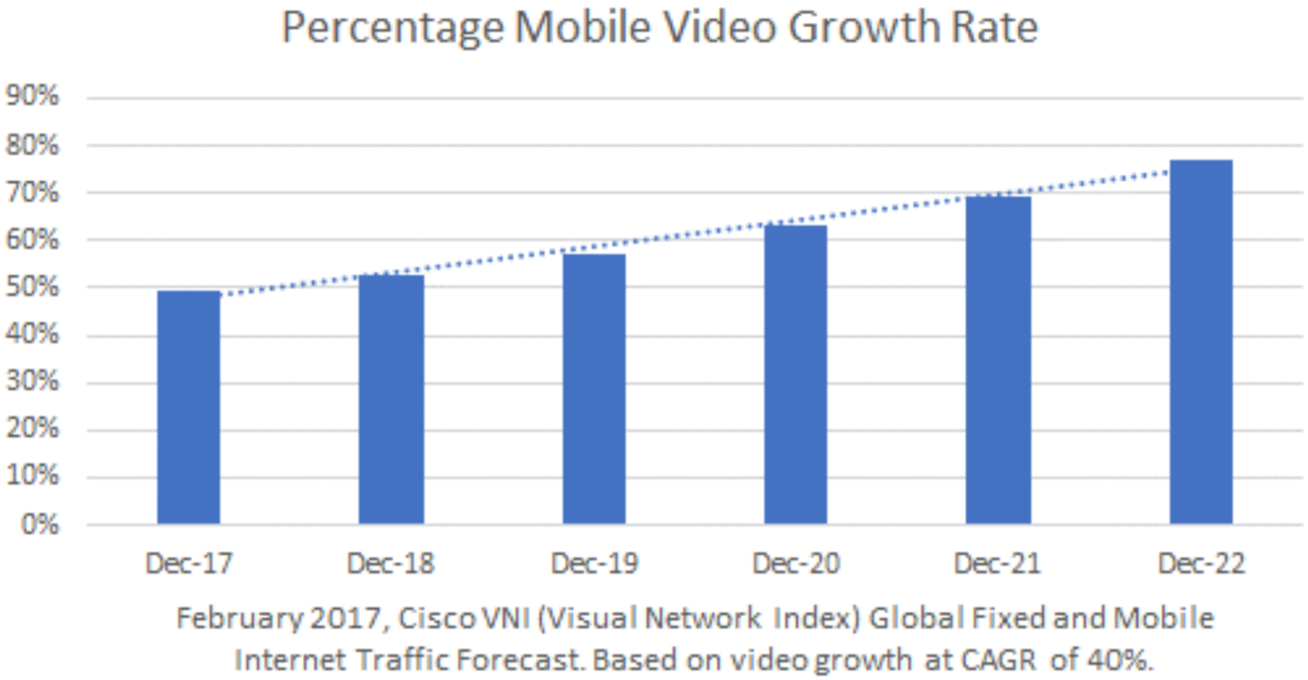


Figure 1: A sample black and white graphic. [?]

NoSQL Databases in support of Big Data and Analytics

Uma M Kugan
Indiana University
711 N Park Avue
Bloomington, IN 47408, USA
umkugan@iu.edu

ABSTRACT

This paper will help us identify how NoSQL is efficient and cost effective in handling big data and also will highlight on why Big Data can't be handled in traditional RDBMS.

KEYWORDS

i523, hid323, NoSQL

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.Organizations that collect large amounts of unstructured data are increasingly turning to nonrelational databases, now frequently called NoSQL databases.[2] 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[8].

2 WHY 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[4]. The term NoSQL means that the database doesn't follow the relational model espoused by E.F Codd in his 1970 paper A Relational Model of Data for Large Shared Data Banks[7] which would become the basis for all modern RDBMS. NoSQL doesn't mean NO to SQL. It means Not Only SQL. NoSQL means storage is just nonvolatile object store with no maintenance concerns. Most NOSQL DB's are open source which allows everyone to evaluate the tool of their choice at low cost.

3 NOSQL TYPES

In [1] Edlich et al. identify four classes of NoSQL systems as 'Core-NoSQL' systems: Key-Value stores, Wide column stores, Graph databases and Document stores.

Key-Value Stores: It is a very basic type of non-relational database where every item (value) is stored as an attribute name (key), with its value. e.g. Redis

Wide Column Stores: Every record in the stores may differ in the number of columns. This is very important factor for analytics because it needs very low I/O and also reduces the volume of data that are read to the disk. e.g. HBase 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. e.g. Neo4J

Document Stores: It stores the data as document typically in Jason or XML format.It is widely used due to its flexibility and ability to query the data. e.g. MongoDB and CouchDB.

4 NOSQL TYPE PERFORMANCE COMPARISON

Ben Scofield rated different categories of NoSQL databases as follows [5]

Data Models	Performance	Scalability	Flexibility	Comp
Column-oriented store	high	high	moderate	lo
Document-oriented store	high	variable (high)	high	lo
Graph database	variable	variable	high	hi
Keyfi?!value store	high	high	high	no
Relational database	variable	variable	low	mod

5 NOSQL FOR BIGDATA

Following factors have to be considered while evaluating NoSQL for Big Data Projects:

5.1 Solution Based on the project Requirements:

Real time Updates for 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.

Publish/Subscribe - NoSQL is the best fit where the enterprise doesn't require complex messaging features for publishing/subscribing.

Document based - Application where data structure is not restricted by schema, NoSQL comes in hand in such places.

5.2 Limitation of traditional Databases:

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 i.e. Atomicity, Consistency, Integrity and Durability and which of course is its strength to process transactional data while the drawback is it can't handle larger volume of data without impacting the performance. If there are use cases where we don't 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't be converted to store in tables, it becomes cumbersome to handle such situations.

6 HOW TO HANDLE RELATIONAL DATA IN NOSQL

NoSQL database in general can't perform joins between data structures and hence the schema has to be designed in such a way so that it can support joins. Below are the key things that needs to be considered to handle relational data in a NoSQL.

6.1 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.

6.2 Denormalize the data :

For faster retrieval of data, it is essential to compromise on denormalizing the data rather than storing only foreign keys.

7 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 DB's we are going to migrate to. [6]

7.1 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's going to be migrated to NoSQL. All the details should be documented well so that we don't have to face unplanned surprises at the end.

7.2 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.

7.3 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.

7.4 Testing

Testing is always very critical and crucial for any migration projects. All aspects of testing from unit testing, functional testing, load testing, integration testing, user acceptance testing etc., have to be carried out and outputs have to be clearly documented.

7.5 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.

8 ADVANTAGES AND DISADVANTAGES 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 don't enforce schema. It is very flexible and let the users to decide to use the data the way they want.

NoSQL databases have the ability to distribute the database across multiple geographic regions to withstand regional failures and enable data localization. Unlike relational databases, NoSQL databases generally have no requirement for separate applications or expensive add-ons to implement replication.[3]

Since NOSQL doesn't enforce atomicity and hence it is not reliable where data accuracy is very critical.RDBMS are much more matured and the best technical support is available.So there is always fear of unknown until the technology gets widely accepted and used.

9 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 collegue Michael Macal for their support and suggestions to write this paper and also to my fellow classmate Andres Castro Benavides for his support.

REFERENCES

[1] S. Edlich, A. Friedland, J. Hampe, and B.Brauer. Oct 2012. NoSQL: Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Hanser Fachbuchverlag. 6 (Oct 2012).

[2] Neal Leavitt. 2010. Will NoSQL Databases Live Up to Their Promise?. (2010). http://www.leavcom.com/pdf/NoSQL.pdf

[3] MongoDB. 2016. *Top 5 Considerations When Evaluating NoSQL Databases*. Technical Report. MongoDB. https://www.mongodb.com/nosql-explained

[4] Editor P. BAXENDALE. June 1970. (June 1970). http:/www.seas.upenn.edu/~zives/03f/cis550/codd.pdf

[5] Ben Scofield. Jan 14 2010. NoSQL - Death to Relational Databases. (Jan 14 2010). http://www.slideshare.net/bscofield/nosql-codemash-2010

[6] Nathaniel Slater. March 2015. Best Practices for Migrating from RDBMS to Amazon DynamoDB- Leverage the Power of NoSQL for Suitable Workloads. (March 2015). https://d0.awsstatic.com/whitepapers/migration-best-practices-rdbms-to-dynamodb.pdf

[7] C. Strozzi. July 2012. Nosql relational database management system. (July 2012). http:/www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/HomePage

[8] Aspire System. 2014. *BigData with NoSQL*. Technical Report. Aspire System. http:/www.aspiresys.com/WhitePapers/BigData_with_NoSQL_Whitepaper.pdf?pdf=nosql-whitepaper

Amazon Web Services (AWS) in Support of Big Data and Analytics

Peter Russell
University of Indiana - Bloomington
petrusse@iu.edu

ABSTRACT

This paper will explore the logistics of Amazon Web Services and how companies are currently utilizing the service to process their big data needs.

KEYWORDS

Big Data, 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)[1]. As a $10 billion a year line of business for Amazon, the revenue stream is incredibly diversified across multiple product offerings. One of these categories, which can broadly be described as 'business analytics,' have helped companies gain new insights into their customer experiences and competitive landscape.

REFERENCES

[1] Synergy Research Group. 2016. AWS Remains Dominant Despite Microsoft and Google Growth Surges. Website. (Feb. 2016).

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

This paper will analyze the processing power of docker with big data use cases

KEYWORDS

i523

1 INTRODUCTION

ACKNOWLEDGMENTS

REFERENCES

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Rono-Hills
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
30 Shuangqing Rd
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

My great Big Dat Paper

YuanMing Huang
Indiana University
800 N Union st
Bloomington, Indiana 47408
huang226@iu.edu

ABSTRACT

THIS IS AN ABSTRACT

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

This is an indtroduction

2 THE BODY OF THE PAPER

this is the body

3 CONCLUSIONS

this is the conclusion

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

What Separates Big Data from Lots of Data

Gabriel Jones
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
gabejone@indiana.edu

ABSTRACT

TIn this paper, we will briefly analyze the history of data to show how having *lots of data* stored in large databases 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 traditional large data storage and analysis. We conclude that organizations that do not understand this paradigmatic shift are more likely to fail in big data projects.

KEYWORDS

i523

1 INTRODUCTION

This is my introduction. [1]

2 CONCLUSIONS

I conclude that...

ACKNOWLEDGMENTS

Generic acknowledgements

REFERENCES

[1] Carl Lagoze. 2014. Big Data, data integrity, and the fracturing of the control zone. *Big Data and Society* (NO 2014). https://doi.org/10.1177/2053951714558281

Big Data and Analytics in Block Chain

Ashok Kuppuraj
Indiana University
Bloomington, Indiana 43017-6221
akuppura@iu.edu

ABSTRACT

This paper describes an idea how Big data and its technologies helps in augmenting or improving the current Block chain technology and overcome one of the problems around it like non-real time transaction time and .

KEYWORDS

Big Data, Block Chain i523

1 INTRODUCTION

The Objective of this project to concur the abilities of the two broad topics in the current technology world, the two B's , Big Data and Block Chain. Block chain and Bigdata both are still a evolving technologies, which gives me enough opportunity to explore and invent new conepts for its own good. As these are still evolving, we can leverage ones solution on the other. To leverage eaach ones problems and solutions, we must first identify the similarities in the two frameworks and how these similarities are related and what solution we are going to choose.

[1].

2 WHAT IS BIG DATA

3 CONCLUSION

Put here an conclusion. Conlcusions and abstracts must not have any citations in the section.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

Big Data Analytics Architecture for Real-Time Traffic Control

Syam Sundar Herle
Indiana University
2965 E Amy Ln
Bloomington, Indiana 47408-4200
syampara@umail.iu.edu

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523, hid219, LaTeX, public tranist, route optimization

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size[**?**].

ACKNOWLEDGMENTS

The authors would like to thank Prof..

REFERENCES

Optimizing Mass Transit Bus Routes with Big Data

Matthew Schwartzer
Indiana University
919 E 10th St
Bloomington, Indiana 43017-6221
mabschwa@indiana.edu

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523, hid225, LaTeX, public tranist, route optimization

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size[1].

ACKNOWLEDGMENTS

The authors would like to thank Prof..

REFERENCES

[1] Keven Richly, Ralf Teusner, Alexander Immer, Fabian Windheuser, and Lennard Wolf. 2015. Optimizing Routes of Public Transportation Systems by Analyzing the Data of Taxi Rides. In *Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics (UrbanGIS'15)*. ACM, New York, NY, USA, 70–76. https://doi.org/10.1145/2835022.2835035

Big Data Applications in Self-Driving Cars

Borga Edionse Usifo
Indiana University Bloomington
107 S Indiana Ave
Bloomington, Indiana 47405
busifo@iu.edu

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Huiyi Chen
Institute for Clarity in Documentation
2451 E. 10TH ST., 612
Bloomington, Indiana 47408
huiychen@indiana.edu

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

This is my Intro

2 THE BODY OF THE PAPER

3 CONCLUSIONS

This is my conclusion.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

REFERENCES

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

G.K.M. Tobin
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Rono-Hills
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
30 Shuangqing Rd
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

ACM proceedings, LaTeX, text tagging

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size.

2 THE BODY OF THE PAPER

Typically, the body of a paper is organized into a hierarchical structure, with numbered or unnumbered headings for sections, subsections, sub-subsections, and even smaller sections. The command `\section` that precedes this paragraph is part of such a hierarchy. LaTeX handles the numbering and placement of these headings for you, when you use the appropriate heading commands around the titles of the headings. If you want a sub-subsection or smaller part to be unnumbered in your output, simply append an asterisk to the command name. Examples of both numbered and unnumbered headings will appear throughout the balance of this sample document.

Because the entire article is contained in the **document** environment, you can indicate the start of a new paragraph with a blank line in your input file; that is why this sentence forms a separate paragraph.

2.1 Type Changes and *Special* Characters

We have already seen several typeface changes in this sample. You can indicate italicized words or phrases in your text with the command `\textit`; emboldening with the command `\textbf` and typewriter-style (for instance, for computer code) with `\texttt`. But remember, you do not have to indicate typestyle changes when such changes are part of the *structural* elements of your article; for instance, the heading of this subsection will be in a sans serif[1] typeface, but that is handled by the document class file. Take care with the use of the curly braces in typeface changes; they mark the beginning and end of the text that is to be in the different typeface.

You can use whatever symbols, accented characters, or non-English characters you need anywhere in your document; you can find a complete list of what is available in the *LaTeX User's Guide* [25].

2.2 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

2.2.1 Inline (In-text) Equations. A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [25]; this section will simply show a few examples of in-text equations in context. Notice how this equation:
$\lim_{n\to\infty} x = 0$,
set here in in-line math style, looks slightly different when set in display style. (See next section).

2.2.2 Display Equations. A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

2.3 Citations

Citations to articles [6–8, 18], conference proceedings [8] or maybe books [25, 33] listed in the Bibliography section of your article will

[1] Another footnote here. Let's make this a rather long one to see how it looks. Footnotes must be avoided.

occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [25]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide* by Lamport [25].

This article shows only the plainest form of the citation command, using `\cite`.

Some examples. A paginated journal article [2], an enumerated journal article [11], a reference to an entire issue [10], a monograph (whole book) [24], a monograph/whole book in a series (see 2a in spec. document) [17], a divisible-book such as an anthology or compilation [13] followed by the same example, however we only output the series if the volume number is given [14] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [36], a chapter in a divisible book in a series [12], a multi-volume work as book [23], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [4], a proceedings article with all possible elements [35], an example of an enumerated proceedings article [15], an informally published work [16], a doctoral dissertation [9], a master's thesis: [5], an online document / world wide web resource [1, 29, 37], a video game (Case 1) [28] and (Case 2) [27] and [26] and (Case 3) a patent [34], work accepted for publication [30], 'YYYYb'-test for prolific author [31] and [32]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [22]. Boris / Barbara Beeton: multi-volume works as books [20] and [19].

A couple of citations with DOIs: [21, 22].

Online citations: [37–39].

We use jabref to manage all citations. A paper without managing a bib file will be returned without review. in the bibtex file all urls are added to rfernces with the *url* filed. They are not to be included in the *howpublished* or *note* field.

2.4 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. ACM uses two types of these constructs: theorem-like and definition-like.

Here is a theorem:

THEOREM 2.1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)\,dt = G(b) - G(a).$$

Here is a definition:

Definition 2.2. If z is irrational, then by e^z we mean the unique number that has logarithm z:

$$\log e^z = z.$$

The pre-defined theorem-like constructs are **theorem**, **conjecture**, **proposition**, **lemma** and **corollary**. The pre-defined definition-like constructs are **example** and **definition**. You can add your own constructs using the *amsthm* interface [3]. The styles used in the `\theoremstyle` command are **acmplain** and **acmdefinition**.

Another construct is **proof**, for example,

PROOF. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

3 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

Generated by bibtex from your `.bib` file. Run latex, then bibtex, then latex twice (to resolve references) to create the `.bbl` file. Insert that `.bbl` file into the `.tex` source file and comment out the command `\thebibliography`.

4 MORE HELP FOR THE HARDY

Of course, reading the source code is always useful. The file `acmart.pdf` contains both the user guide and the commented code.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the matlab code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientsts' Support Program (http://www.nnsf.cn/youngscientsts).

REFERENCES

[1] Rafal Ablamowicz and Bertfried Fauser. 2007. CLIFFORD: a Maple 11 Package for Clifford Algebra Computations, version 11. (2007). Retrieved February 28, 2008 from http://math.tntech.edu/rafal/cliff11/index.html

[2] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. https://doi.org/10.1145/1188913.1188915

[3] American Mathematical Society 2015. *Using the amsthm Package*. American Mathematical Society. http://www.ctan.org/pkg/amsthm

[4] Sten Andler. 1979. Predicate Path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226–236. https://doi.org/10.1145/567752.567774

[5] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[6] Mic Bowman, Saumya K. Debray, and Larry L. Peterson. 1993. Reasoning About Naming Systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 795–825. https://doi.org/10.1145/161468.161471

[7] Johannes Braams. 1991. Babel, a Multilingual Style-Option System for Use with LaTeX's Standard Document Styles. *TUGboat* 12, 2 (June 1991), 291–301.

[8] Malcolm Clark. 1991. Post Congress Tristesse. In *TeX90 Conference Proceedings*. TeX Users Group, 84–89.

[9] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[10] Jacques Cohen (Ed.). 1996. Special issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[11] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[12] Bruce P. Douglass, David Harel, and Mark B. Trakhtenbrot. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. https://doi.org/10.1007/3-540-65193-4_29

[13] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

[14] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. https://doi.org/10.1007/3-540-09237-4

[15] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07)*. USENIX Association, Berkley, CA, Article 7, 9 pages.

[16] David Harel. 1978. *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER*. MIT Research Lab Technical Report TR-200. Massachusetts Institute of Technology, Cambridge, MA.

[17] David Harel. 1979. *First-Order Dynamic Logic*. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. https://doi.org/10.1007/3-540-09237-4

[18] Maurice Herlihy. 1993. A Methodology for Implementing Highly Concurrent Data Objects. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 745–770. https://doi.org/10.1145/161468.161469

[19] Lars Hörmander. 1985. *The analysis of linear partial differential operators. III*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. viii+525 pages. Pseudodifferential operators.

[20] Lars Hörmander. 1985. *The analysis of linear partial differential operators. IV*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. vii+352 pages. Fourier integral operators.

[21] IEEE 2004. IEEE TCSC Executive Committee. In *Proceedings of the IEEE International Conference on Web Services (ICWS '04)*. IEEE Computer Society, Washington, DC, USA, 21–22. https://doi.org/10.1109/ICWS.2004.64

[22] Markus Kirschmer and John Voight. 2010. Algorithmic Enumeration of Ideal Classes for Quaternion Orders. *SIAM J. Comput.* 39, 5 (Jan. 2010), 1714–1747. https://doi.org/10.1137/080734467

[23] Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.)*. Addison Wesley Longman Publishing Co., Inc.

[24] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

[25] Leslie Lamport. 1986. *LaTeX: A Document Preparation System*. Addison-Wesley, Reading, MA.

[26] Newton Lee. 2005. Interview with Bill Kinder: January 13, 2005. Video. *Comput. Entertain.* 3, 1, Article 4 (Jan.-March 2005). https://doi.org/10.1145/1057270.1057278

[27] Dave Novak. 2003. Solder man. Video. In *ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003)*. ACM Press, New York, NY, 4. https://doi.org/99.9999/woot07-S422

[28] Barack Obama. 2008. A more perfect union. Video. (5 March 2008). Retrieved March 21, 2008 from http://video.google.com/videoplay?docid=6528042696351994555

[29] Poker-Edge.Com. 2006. Stats and Analysis. (March 2006). Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php

[30] Bernard Rous. 2008. The Enabling of Digital Libraries. *Digital Libraries* 12, 3, Article 5 (July 2008). To appear.

[31] Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

[32] Mehdi Saeedi, Morteza Saheb Zamani, Mehdi Sedighi, and Zahra Sasanian. 2010. Synthesis of Reversible Circuit Using Cycle-Based Approach. *J. Emerg. Technol. Comput. Syst.* 6, 4 (Dec. 2010).

[33] S.L. Salas and Einar Hille. 1978. *Calculus: One and Several Variable*. John Wiley and Sons, New York.

[34] Joseph Scientist. 2009. The fountain of youth. (Aug. 2009). Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.

[35] Stan W. Smith. 2010. An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and Computers (LAC '10)*, Reginald N. Smythe and Alexander Noble (Eds.), Vol. 3. Paparazzi Press, Milan Italy, 422–431. https://doi.org/99.9999/woot07-S422

[36] Asad Z. Spector. 1990. Achieving application requirements. In *Distributed Systems* (2nd. ed.), Sape Mullender (Ed.). ACM Press, New York, NY, 19–33. https://doi.org/10.1145/90417.90738

[37] Harry Thornburg. 2001. Introduction to Bayesian Statistics. (March 2001). Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

[38] TUG 2017. Institutional members of the TeX Users Group. (2017). Retrieved May 27, 2017 from http://wwtug.org/instmem.html

[39] Boris Veytsman. [n. d.]. acmart—Class for typesetting publications of ACM. ([n. d.]). Retrieved May 27, 2017 from http://www.ctan.org/pkg/acmart

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

The Internet of Things and Big Data

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, Data Science, Internet of Things, IoT, Smart Devices, Sensors, Actuators, Big Data Analytics, Cloud Computing

1 INTRODUCTION

The Internet of Things (*IoT*) is the network of physical devices, vehicles, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to collect and exchange data[6]. 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. Today, over 50% of IoT activity is centered in manufacturing, transportation, smart environments and consumer applications like wearable gadgets, but within five years all industries will have rolled out IoT initiatives. *Gartner, Inc.* forecasts that 8.4 billion connected things will be in use worldwide by end of 2017 and will reach 20.4 billion by 2020[2].

2 IOT INTUITION

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*. 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:

Smartphones - With smartphone's range of sensors (accelerometer, gyro, video, proximity, compass, GPS, etc.) and connectivity options (cell, wi-fi, bluetooth, etc.) user has well equipped IoT device that can automatically monitor movements, location and workouts throughout the day.

Smart Homes - Here is an example of smart home enabled by IoT devices: The user arrives home and his car communicates with the garage to open the door. The thermostat is automatically adjusted to his preferred temperature due to sensing his proximity. He walks through his door as it unlocks in response to his smartphone or RFID implant. The home lighting is smartly turned on at dark.

Smart Cities - Smart surveillance, safer and automated transportation, smarter energy management systems and environmental monitoring are all examples of IoT applications for smart cities.

Smart Medical Alerts - The Proteus ingestible pill sensor is powered by contact with stomach fluid and communicates a signal that determines the timing of when patient took her medication and the identity of the pill. This information is transferred to a patch worn on the skin to be logged for the patient and her doctor's reference. Heart rate, body position and activity can also be detected accordingly.

Smart Aircrafts - Rolls-Royce is using Azure Cloud Stream Analytics and Power Business Intelligence (BI) to link up sensor data from its engines with more contextual information like air traffic control, route data, weather and fuel usage to get a complete report of the health of its aircraft engines.

3 ALLIANCE WITH BIG DATA

The true value of IoT is not in the internet connected devices themselves; the value lies in making context-aware relevant data and converting the result into enterprise-grade, tangible and *actionable* business insights. The IoT and Big Data are intimately connected: billions of internet-connected things will, by definition, generate massive amounts of data. As the *things* turn more digital, IoT will analyze complex data structures and respond intelligently in real time.

Big Data, meanwhile, is characterized by *four Vs* - volume, variety, velocity and veracity[5]. That is, data come in large amounts (*volume*), with a combination of structured and unstructured data (*variety*), arrive at often real-time speed (*velocity*) and can be of uncertain source (*veracity*). Such information is unsuitable for processing using traditional SQL-queried relational database management systems (RDBMSs), which is why a cluster of alternative tools – notably Apache's open-source *Hadoop* distributed data processing system, various *NoSQL* databases and a range of business intelligence (*BI*) platforms - have evolved to serve such a complex data process.

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:

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. *Sensors* measure physical inputs and transform them into raw data; *actuators* act on the signal from the sensors and convert it into output, which is then digitally storable for access and analysis. These tiny innovations can measure anything ranging from temperature, force, flow, position to even light intensity then can be attached to everything from smartphones to medical devices and then record & send data onto the cloud[3].

Network Connectivity in the devices is achieved through: wireless/wired, wi-fi, bluetooth, zigbee, VPN and cellular - 2G/3G/LTE/4G. Thread technology is emerging as an alternative for home automation applications and Whitespace TV technologies being implemented in major cities for wider area IoT-based use cases. Depending on the application, factors such as range, data requirements, security, power demands and battery life will dictate the choice of one or some form of combination of the technologies. In March 2015, the Internet Architecture Board - a group within the Internet Society that oversees the technical evolution of the internet - released a guide to IoT networking. This outlined four common communication models used by IoT smart devices: Device-to-Device, Device-to-Cloud, Device-to-Gateway, and Back-End Data-Sharing[4].

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 the 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 within which devices and applications can exchange data securely over wired or wireless networks. To address these challenges and for better IoT interoperability, one key player, *OneM2M* issued Release 1, a set of 10 specifications covering requirements, architecture, Application Programming Interface (API) specifications, security solutions and mapping them to common industry protocols[1].

Cloud and Big Data Analytics - The cloud 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.

Data Analytics involves statistical tools and techniques with business acumen to bring out hidden information from the data. Advanced types of data analytics include data mining, which involves sorting through large data sets to identify trends, patterns and relationships; predictive analytics, which seeks to predict customer behavior, equipment failures and other future events; and machine learning, an artificial intelligence technique that uses automated algorithms to churn through data sets more quickly than data scientists can do via conventional analytical modeling. Text mining provides a means of analyzing documents, emails and other text-based content. Big Data analytics applies data mining, predictive analytics and machine learning tools to volume of data coming from various sources with various types of data formats.

Big Data analytics, in the context of the IoT, refers to sensor analog inputs being converted to digital data, analyzed, and resulting in a response going back to the device. Much of this data is in an unstructured form, making it difficult to put into structured tables with rows and columns. To extract valuable information from this complex data, Big Data applications often rely on cutting edge analytics involving data science. Distributed computers in the cloud running sophisticated algorithms can help enhance the veracity of information by data mining through the noise created by the massive volume, variety, and velocity. Some analytics may need to be performed using edge or mesh computing, some in the data center and some in a cloud environment, depending on the trade-off of speed versus depth. IoT analytics applications can help companies understand the IoT data at their disposal, with an eye toward reducing maintenance costs, avoiding equipment failures and improving business operations.

5 CONCLUSION

Internet of Things shaping human life with greater connectivity and ultimate functionality, and all this is happening through ubiquitous networking to the Internet. There is seemingly no limit to what can be connected to the Internet. IoT will become more personal and predictive. 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 need with technologies, data, process, insight, action and people. The *T* of IoT is clearly important, but too often, it is the only area of focus when examining IoT in business. Rest of the systems need to be instrumented to leverage the data: communicating it to the right place for action - whether the cloud, data center, or edge - and then using analytics to understand data patterns and craft a response to fix or optimize. However, security and privacy will be the top considerations for companies developing IoT devices. Innovative organizations are starting to put this to use today.

REFERENCES

[1] 2015. IoT Interoperability. (Jan. 2015). http://www.onem2m.org/images/files/oneM2M-whitepaper-January-2015.pdf

[2] 2017. Garner Press Release. (Feb. 2017). http://www.gartner.com/newsroom/id/3598917

[3] Hakim Cassimally Adrian McEwen. 2014. *Designing the Internet of Things.* Wiley.

[4] Lyman Chapin Karen Rose, Scott Eldridge. 2015. *The Internet of Things: An Overview.* Technical Report. https://www.internetsociety.org/resources/doc/2015/iot-overview

[5] Wikipedia. 2017. Big Data. (2017). https://en.wikipedia.org/wiki/Big_data [Online; accessed 23-Sept-2017].

[6] Wikipedia. 2017. Internet of things. (2017). https://en.wikipedia.org/wiki/Internet_of_things [Online; accessed 23-Sept-2017].

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

My great Big Dat Paper

Ben Trovato
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
trovato@corporation.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.

KEYWORDS

i523

1 INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes, a specified live area, centered on the page, specified size of margins, specified column width and gutter size [1].

ACKNOWLEDGMENTS

The authors would like to thank

REFERENCES

[1] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4