EXPLORING BIGDATA

Touro College Student Chapter
of the
Association for Computing Machinery

LANDER COLLEGE
OF ARTS & SCIENCES
A DIVISION OF TOURO COLLEGE IN FLATBUSH
Where Knowledge and Values Meet

TABLE OF CONTENTS

EDITORIAL

Welcome to The ACM Student Chapter Publication!

We are very excited to present this year's publication, featuring articles exploring Big Data and the power it wields. Big Data is at the forefront of technological innovation, and it is our mission to demystify this field of study.

Before we can begin, it is important to acknowledge all the hard work that went into making this publication possible.

First and foremost, we would like to express our appreciation to Dr. Fink, Deputy Chair of the Computer Science Department.

We also thank the writing department for editing the articles.

Most importantly, we would like to recognize those who contributed to the publication, dedicating hours of time researching, writing, and editing the articles contained herein.

We hope you enjoy this publication!

Editors-in-Chief,
Rina Younger and Hudi Teitelbaum

Welcome to the Touro College Student Chapter of the ACM Newsletter

Welcome to the Spring 2020 edition of the Touro College Student Chapter of the ACM (Association of Computing Machinery) newsletter published by the Flatbush Campus Division. The purpose of the student chapter of the ACM, as outlined in its bylaws, is "to promote an increased knowledge of and greater interest in the science, design, development, construction, languages, management, and applications of modern computing." This newsletter certainly helps to accomplish that goal.

The authors of the articles in this edition of our newsletter deserve special credit. They completed their articles and put the newsletter together under adverse conditions and in difficult times. Their perseverance is truly admirable.

Big data, the theme of this newsletter, is a fascinating topic, which has evolved significantly in the past few years, and will, undoubtedly, continue to develop and play an increasingly more central role in the field of computer science in the years to come. As the quantity of data being produced annually continues to increase exponentially, so do the challenges inherent in analyzing, interpreting, and utilizing that data. Reading the articles in this newsletter, as well as taking advantage of other opportunities to learn about big data, will certainly prove to be beneficial.

A glimpse of the relevance and currency of big data can be seen from the HackerRank Tech Recruiting Benchmark Report, which was derived from a recent survey of thousands of tech hiring leaders. In portraying the tech recruiting landscape, the report identified big data and analytics as the number one driver of tech hiring. 50.4% of respondents selected big data from a list of ten initiatives they felt were driving technical hiring at their company. The report noted that this is likely due, at least in part, to the growth of big data and analytics based solutions, which are slated to see a collective annual growth rate of 13% through 2022, according to International Data Corporation (IDC).

As you will see from the articles in this newsletter, big data is closely related to many of the other most popular areas of computer science. Topics such as cloud computing, machine learning, security and privacy, natural language processing, and other state-of-the-art topics are addressed in conjunction with Big Data. There is no question that big data will remain a valuable and even critical component of computer science for many years to come.

As evidenced by this publication, our student chapter of the ACM yields insight into computing as a science and a profession. Various activities, such as doing research, writing articles, and attending meetings, can help keep you informed of state-of-the-art developments. Participation in a professional organization beyond the required course work is evidence of serious interest and dedication to advancing in the technology industry, something employers are always looking for. Keep in mind, however, that you may be asked on an interview what you did as a member of the Student Chapter of the ACM. The more actively you have participated in chapter events and publications, the more impressive your answer will be.

I recommend you consider writing an article for the next edition of the newsletter. Doing so will

give you the opportunity to participate in a meaningful way, while at the same time research and learn about a cutting-edge facet of technology. Please approach me, or one of the officers, if you would like to volunteer or if you have any ideas for future chapter speakers or events.

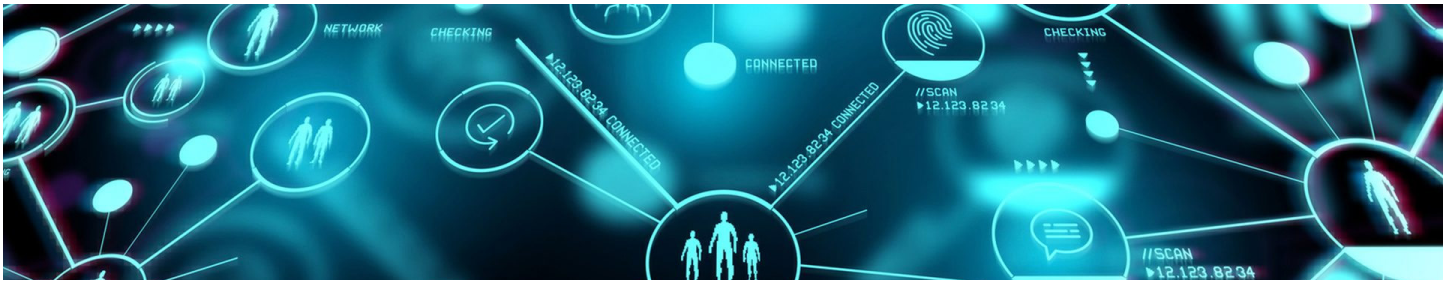
Our current officers are: Rina Younger – President, Yehudis Teitelbaum – Vice President, Nechama Leah Perlman – Treasurer, and Ita Goldfarb – Secretary. We will be appointing some new officers next year to replace those who are graduating. Please let me know if you would like to be considered for one of the available positions. As mentioned above, serving as an officer is a great way to enhance your resume with leadership skills.

I would like to thank Yehudis Teitelbaum for all of her hard work on the formatting and layout of this newsletter. Of course, we thank Nechama Perlman, Chaya Lev, Bas Tziyon Friedler, Miriam Kamin, Yehudis Teitelbaum, Sarah Weiss, Rina Younger, Chana Fine, Kara Noble, Aaron Farntrog, Chaya Sara Zitwer, Penina Ziegler, Tehila Raful, Hanah Lavian, and Ari Weinberg for voluntarily writing the articles in this edition of our newsletter.

As we approach the end of another academic year, I am proud to say that current Touro College computer science students and graduates from the past 12 months have landed jobs and internships in Americans Software Resources, Avant-garde Health, B&H Photo & Electronics, Blueswitch, Central Analysis Bureau, Crawford Lake Capital Management, Cross River Bank, Federal Reserve Bank, Financial Industry Regulatory Authority, General Atomics, Google, Grey Matter, Guggenheim Partners, Harris Corporation, Healthfirst, Infobase Publishing, Intersoft Solutions, Kew Systems, Litify, Megadata Health Systems, NBC Universal Media, New York Life Insurance Company, Northrop Grumman, NYC Department of Education, Premier Data Group, Reliable Health Systems, Rockit Solutions, Stringbean Technologies, Taryag Analytics, UBS Group, Upliftr, YeshivaNet, and many other companies. Feedback from these graduates and their employers has been that they are well prepared for these positions.

As always, I would like to conclude by expressing my appreciation to Dr. Issac Herskowitz for the phenomenal job he has done leading and guiding the department and the faculty of the Department of Computer Science for the extraordinary job they have done preparing our students for this ever-changing career. They expend countless hours learning new technologies, revising the curriculum, and developing new approaches to ensure that our graduates will have state-of-the art skills and be best prepared for the ever-changing careers in technology they seek. Without Dean Herskowitz and our faculty's hard work and willingness to go the extra mile, our graduates would not be where they are today.

Best wishes for continued success,
Shmuel Fink
Deputy Chair, Department of Computer Science
Touro College



WHAT IS BIG DATA?

BY: NECHAMA PERLMAN

How many times a day do you sign up to a website and are prompted to enter your personal information? How many times a day do you order something online, and again are prompted to enter your information? If you're a regular person living in the 21st century, the answer to both these questions is - many, many times a day. Every time you enter your information to a website it is stored in a database. Websites aren't just for your community. Websites are on the World Wide Web, meaning they're available to anyone. If a website is available to anyone, then the website needs to have a database big enough to hold the information of all of its clients. This is where big data comes into play.

The term "Big Data" can be used two different ways. The first way refers to a methodology of handling data sets that are too large or complex for traditional data-processing application software.

The term big data can also be used to describe data. Big data doesn't necessarily refer to a specific type of file, like a jpg or png. It can refer to a file that is too large for the server to compute in a regular amount of time (1). For example, take a company like Amazon. Amazon Prime has over 100 million members according to businessinsiders.net (2), which means 100 million usernames, 100 million passwords, 100 million email addresses etc. All this information needs to be stored in a database that allows Amazon to easily access it. For companies like Amazon, the amount of data they are dealing with is too large for a common database program. They aren't dealing with regular data, they are dealing with big data. Therefore they must use big data techniques to manage this data .

Big data has a larger role today than it did in previous times because of the growth of technology. Between our phones, smart watches, cars, electronic doorbells, etc., data is constantly being generated and recorded. All of this information can be a good thing. It is very convenient when you plug your phone into Apple CarPlay and it already knows that you're going to work, because you go there every day at the same time. It is also convenient when your phone gives you suggestions of what to do, like to text your friend, because you usually text her at that time. These things can be helpful, but it is important to remember that everything we do is being recorded. The government has access to all this information, and they can use it to track your daily movements or gain information for research they conduct. However, as long as you're not doing anything suspicious or illegal you have nothing to fear.

You may be wondering what big data means for you in your day to day life. For the average college student, big data won't directly play a big role in your life. However, for White-collar professionals, big data will impact life greatly. On the positive side, because big data can hold and process that much more data, there will be more information available to help professionals do their jobs. For example, big data can give doctors all around the world access to informa-

tion, which will improve how they choose to treat a patient. On the other hand, according to author Kenneth Cukier, along with all the information coming in and being held by big data, there is increasing knowledge of artificial intelligence and machine learning, which may make some of these white-collar jobs obsolete (3).

Big data is the result of society's technological advancements, and will continue to play a large role in future advancements. Big data enables us to store exponentially larger amounts of data , which will give us more information to study in order to advance and better our society. Big data will undoubtedly be a key component in the technology that will shape our future.

WORKS CITED

(1) Ner. What Is Big Data. NetVercity, 2014. https://www.youtube.com/watch?v=tkOwlXUaGMM.

(2) Green, Dennis. "A Survey Found That Amazon Prime Membership Is Soaring to New Heights - but One Trend Should Worry the Company." Business Insider. Business Insider, January 18, 2019. https://www.businessinsider.com/amazon-more-than-100-million-prime-members-us-survey-2019-1.

(3) Cukier, Kenneth. TED. Accessed March 15, 2020. https://www.ted.com/talks/kenneth_cukier_big_data_is_better_data?language=en#t-6876.

DATA VISUALIZATION AND ITS IMPORTANCE

BY: CHAYA LEV

They say, "a picture is worth a thousand words." With data visualization, that is especially true. Data visualization is the representation of data in a pictorial or graphical format and is used in data analysis to identify patterns and trends. Because of the way the human brain processes information, seeing data visually makes it easier to pinpoint trends and can help detect patterns and correlations that would otherwise go undetected (1). According to Forbes, over 65% of the population are visual learners, so it's easy to understand why looking at a neatly formatted graph or diagram is easier to analyze than looking over thousands of rows on a spreadsheet (2). Presenting data visually also enables concepts to be easily displayed in a more universal manner, since the graphics make the message less dependent on written language. With data visualization, areas that need attention can get recognized, light can be shed onto factors that influence customer behavior and sales volumes can be predicted.

With the rise of big data, data visualization plays a crucial role. Businesses have extensive collections of data that are growing every year. Because of this, they need an efficient way to get valuable insights from their data. Data visualization provides a way for decision makers to be able to access, evaluate and act on data in real-time. Big data visualization goes beyond the classic visualization forms like graphs and pie charts and involves more complex representations like heat maps and fever charts. In a heat map, individual values contained in a matrix are represented as colors. A warm-to-cool color spectrum is used to indicate different values, such as which areas of a webpage receive the most clicks or what the most densely populated areas are on a map. What makes heat maps so popular is the fact that they are a lot more visual than standard visualization forms, and as a result, don't require interpretation (3). A fever chart, on the other hand, is a graphical representation of changing data over time. Fever charts are useful to track changes in a variable over time, whether in business, stock prices, or weather.

Big data visualization requires powerful computer systems to accumulate raw data and process it in order to generate graphical representations that humans can easily comprehend (4). While big data visualization is extremely powerful, there are some unique challenges that come along with this power. One challenge is finding enough visualization specialists who know how to use the right data visualization techniques to ensure that the most is made of the data. Another challenge is ensuring that the right hardware resources are available since big data visualization may require powerful hardware. And lastly, people and processes need to be put into place in order to maintain the data quality so that only accurate data is visualized (5).

Despite these challenges, data visualization is a powerful tool in presenting, understanding and analyzing data. When confronted with the need to present enormous amounts of data, data visualization can be an invaluable way to make the data available to

the widest group or audience. Data visualization provides the tools necessary to go big with big data.

WORKS CITED

(1) "Data Visualization: What It Is and Why It Matters." SAS. Accessed March 15, 2020. https://www.sas.com/en_us/insights/big-data/data-visualization.html#dmtechnical.

(2) McCue, TJ. "Why Infographics Rule." Forbes. Forbes Magazine, April 2, 2013. https://www.forbes.com/sites/tjmccue/2013/01/08/what-is-an-infographic-and-ways-to-make-it-go-viral/#31b825077272.

(3) Hall, Sharon Hurley HallSharon Hurley. "What Is a Heat Map, How to Generate One, Example and Case Studies." The Daily Egg, May 15, 2019. https://www.crazyegg.com/blog/understanding-using-heatmaps-studies/.

(4) Rouse, Margaret. "What Is Data Visualization and Why Is It Important?" SearchBusinessAnalytics. TechTarget, February 20, 2020. https://searchbusinessanalytics.techtarget.com/definition/data-visualization.

(5) "Big Data Visualization." What is Big Data Visualization? Accessed March 15, 2020. https://www.datamation.com/big-data/big-data-visualization.html.

WHAT IS DATA MINING?

BY: BAS TZIYON FRIEDLER

What is data mining? Data mining is, "the process of discovering actionable information from large sets of data….using mathematical analysis to derive patterns and trends that exist in data" (1). In other words, data mining is the analysis of information gathered legally online to predict patterns in order to benefit businesses (2). Data mining is exactly what it sounds like- mining and sifting through collected data. The 'gold' that remains in the sifter is patterns and trends that can be used to predict future events. Data mining companies collect masses of data that is stored in servers or a cloud. They then analyze the data and present it in a way that can be useful. In the words of Jean-Francois Belisle, a marketing and performance director at K3 Media, "A data miner is like the magician Criss Angel that will make appear from your messy ocean of data, insights that will be valuable to your company and may give you a competitive advantage compared to your competitors."(2)

How is this data used in a useful way? After the data is analyzed it can be used in many ways. One way is for "forecasting." The results of the analysis can be used to estimate sales or server downtime (1). For example, a UCLA professor gave an example of a supermarket that used data mining to find out that men who buy diapers on Thursdays and Saturdays were also likely to buy beer. The supermarket can use this information to their benefit and place the beer closer to the diapers, or make sure to sell beer at full price on those days in order to maximize their profit (3). Another way data mining is used is to find out the probability of a consumer buying a product and therefore targeting specific customers (1). For example, when you buy something online, you'll probably notice that you will have ads for that item or similar items popping up on your screen (2). This is because data miners have found a correlation between the product you purchased and the likelihood that you will buy it, or a similar product, again. Data mining can also be used to analyze the items in your online shopping cart and can predict the "next likely event" (1). For example, Amazon uses a technology that analyzes what you buy and recommends similar products to you. Investopedia estimated that over one third of Amazon's sales come from using this technique since many people impulsively buy the products recommended to them (1). Some data mining companies just buy data, analyze it, and create consumer profiles from the data that they then sell (2). These are just some of the many ways that companies use data mining.

How does data mining work? First, the data must be collected. This is done in numerous ways. Many times our personal information is being shared without us even knowing it. Many credit card companies track each time you swipe your card and store information about the purchase (2). Data Miners also get their information from social media. Facebook gives over information including past and current friends, every ad you clicked on, and personal information such as age, gender, name, and place of residence. The New York Times reported that in 2018, Facebook showed the personal information of

users to over 150 technology companies such as Netflix, Microsoft, and Amazon (1). Data Miners also get their information from apps that are downloaded with hidden trackers that send information (1). In a 2019 study by Jama Network Open, nine out of ten apps designed to help stop depression and smoking shared users personal information. And out of those nine companies, only three of them warned their users beforehand. Whether or not we know it, our information is constantly being collected.

Once they have information, Data Miners analyze it using mathematical algorithms. This is done with the help of special data mining software, many of which are semi-automated with already built in algorithms (3). For example, Microsoft has a data mining software known as the Microsoft SQL Server. Data Miners use this program and go through the data mining process. First, the problem or what they are looking to find is defined. Step two is to sift through the collected data and "clean it up"-decide which parts of it to use or not use. Then they explore the data by going through it and finding means, minimums, and maximums, and figuring out standard deviations. If the deviation deviates a lot from the norm, this tells them that either something is wrong with the data or that there is not enough data to work with. Step four is to create a mining model, applying specific math algorithms to the model. The data is then passed through the model and analytical information is produced. Lastly, the results are deployed in real life. Then, Data from these results can be collected and the process can start again (1).

In this century, Data Mining is probably the most effective way to boost companies' sales. With so much data available, much can be done to predict and presume people's behavior. Although it may be sneaky and the thought of it can make one uncomfortable, data mining is not illegal as of yet. In this age of technology, almost every move you make is being tracked and stored somewhere where it can be analyzed and used to benefit companies, and likely disadvantage you.

WORKS CITED

(1) Minewiskan. "Data Mining Concepts." Microsoft Docs, 9 Jan. 2019, docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions.

(2) Turner, Terry. "Data Mining: Consumer Risks & How to Protect Your Information." ConsumerNotice.org, 10 Jan. 2020, www.consumernotice.org/data-protection/mining/.

(3) Brooks, Chad. "What Is Data Mining?" Business News Daily, Businessnewsdaily.com, 18 Feb. 2014, www.businessnewsdaily.com/5947-data-mining.html.

DATA MINING TECHNIQUES

BY: MIRIAM KAMIN

According to Jiawei Han, professor of data mining and director of INARC in Illinois, the incredible amounts of data we have access to is not enough. He says, "we are data-rich, but information poor"(1). Although there is an endless amount of data, in order to utilize that data one must go through a grueling process to make that data accessible, readable, accurate and useful. The process of preparing big data for analysis is referred to as *data mining*.

When organizing big data for analysis, data will likely be incomplete (lacking other necessary data), noisy (containing errors and outlier values) and inconsistent (discrepancies in the categorization of data) (1). Welcome to the real world!

In order to make use of the many methods of data mining, data must be clean of the above issues. Dirty data will produce unreliable outputs and many errors. Users cannot rely on the results of dirty data.

Data must be preprocessed, prior to being mined. Preprocessing is a multi-step process made to clean, integrate, transform and reduce data into simpler, more accessible forms.

The first stage of preprocessing data is cleaning data. Cleaning data means ridding data of inaccuracies, filling in missing values, identifying outliers, and correcting inconsistencies. Data integration is the next step. When combining data from multiple sources, data must be compiled into a common system. Taking data from many sources, as is the case with big data, poses difficulties, as each unique classification system must be reformatted to interact with other unique systems. Again, ridding redundancies is a necessity here, as well as resolving conflicting value systems, such as metric vs. imperial, or taxes in different states causing prices to vary. Next, it is often necessary to transform data to more readable formats and/or scales. Data reduction techniques will be your last step in cleaning your data. When dealing with big data, most often, you'll be dealing with *too much* data. While some of it is necessary, large amounts of information are unnecessary for analysis and should be removed. To further reduce space, data should be consolidated to simpler forms, while still fulfilling its structural needs (1).

Once data is preprocessed, it should be void of critical inaccuracies, redundancies, missing data, and discrepancies. At this point, the data is still not in a readable form. The data must be mined so proper analysis can take place.

There are many different data mining techniques. One of the most basic techniques in data mining is tracking patterns. This means looking for certain changes, and noticing an ebb and flow of values over time. Perhaps holiday seasons, weather, or political changes may produce patterns to follow and analyze. Classification, a more complex technique, forces you to collect various attributes together into categories, which you can then use to draw conclusions and learn more about these attributes. Simi-

lar to classification, association is a technique used to look for specific events or attributes that are correlated to other events or attributes. Perhaps when a certain ideology is popular, certain political circumstances change as well, or when sales of one item rise, sales increase elsewhere as well. Detecting outliers will help point out anomalies that disrupt your understanding of the data set. Investigating that outlier will give you insight on what drove that burst of change, to try replacing it or better service it. Predicting what data you're likely to receive in the future and analyzing future trends can give an idea of what is soon to come. Decision trees are a specific method of modeling data. They help illuminate how inputs affect outputs. When various decision tree models are combined, called a *decision forest,* they can provide deeper insight into data, although are often quite difficult to understand. Statistical organization, as well as creating visualized models, are both common methods used to better express information (2).

Data that is mined and processed must be stored in database management systems to be analyzed by others at a later time. This can be done in the cloud as well.

Another complementary method to data mining is *machine learning*, where the machine will analyze the data by automatically learning the parameters from the data. The machine will use self-learned algorithms to improve its performance over time. Machine learning can send feedback in near real-time. In general, the larger the dataset, the greater the accuracy and performance of machine learning. The machine can determine relationships within the data, and adapt to new norms, without needing to reprogram baselines or key indicators (3).

When mining through big data, you make the most out of the masses of data gathered. With properly set and organized data, you can apply the correct logic, ask the right questions, and recognize true patterns that will revolutionize your data into useful resources of information.

WORKS CITED

(1) Han, Jiawei, and Micheline Kamber. Data Mining, Southeast Asia Edition. Vol. 2nd ed, Morgan Kaufmann, 2006.

(2) Alton, Larry. "The 7 Most Important Data Mining Techniques." Data Science Central, 22 Dec. 2017.

(3) McDaniels, Stacy. "16 Data Mining Techniques: The Complete List." Talend, 29 July 2019.

MEDICAL DATA MINING

BY: YEHUDIS TEITELBAUM

According to the National Center for Health Statistics (NCHS), there were about 883.7 million doctors' visits in 2016 (1). That's at least 883.7 million medical data entries. And what do they do with all the information intake? Every visit is recorded, and symptoms, conclusions, and findings are documented. This information is extremely valuable to medical researchers. Doctors are constantly exploring new diseases, trying to learn how fatal they are, which medications work, how people react to the medications, and what they can do to prevent the diseases. When dealing with myriads of data, it can be hard for even a team of researchers to gather the desired information and patterns. That's where computers and big data analysis comes in. Using a technique called Data Mining, researchers can effectively and efficiently find trends in the data, helping doctors reach important conclusions and perfect diagnosing and treatments.

What is Medical Data Mining? Just as the term "mining" colloquially refers to the extraction of precious materials from mounds of earth, so too, data mining is the action of gleaning meaningful information from an abyss of data. The process of extracting data begins with moving the information to a data warehouse. Here, they won't mess up the original copy and the information is in one place for easier querying of the data. The next step is to clean and organize the data, a process in which errors are singled out and removed, and the data is unified to a single format. Now that the data is properly set up, it is ready to be mined and analyzed (2).

The article, Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse, explains the medical importance of analyzing big data: "Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management." (2) By using Big Data Analytics to track and compare symptoms and treatments, doctors gain insight to the best route of treatment, and faster, more precise diagnosing (3). For example, big data analysis has made diagnosing cancer easier and more accurate. Since approximately 20% of cancer is undetected in its initial stages, where treatment would be the most effective, using the power of AI prediction can save lives (4). By analyzing the vast amount of data and using Artificial intelligence to learn from the analysis, these predictions are made possible. As better techniques are discovered for data mining and data analysis, these predictions are becoming more accurate. In fact, Lunit, a data analytic tool for predicting and diagnosing cancer, is 97-99% accurate. It studies annotated datasets and bases its predictions off them. Because of data imperfections, i.e. trivialities and redundancies, Lunit is not always able to determine a diagnosis. In such cases, Lunit locates the questionable data and passes it on to be reviewed by a doctor. Lunit then reviews the doctor's diagnosis, thereby gaining insight on how to make similar predictions in the future. Once they clean up the data, Lunit is left with the job of analyzing and making predictions based on the vast amount of data at its disposal (5).

There are various approaches to diagnosis prediction from big data. Notably, there are three main prediction algorithms used in medical data analysis: Naive Bayes, Decision List, and KNN.

Naive Bayes uses simple probability statistics to classify the data, create a frequency of occurrence, and make a prediction of likelihood based on it. For example, using attributes such as age, sex, chest pain type, and fasting blood sugar creates all of the possible combinations between these variables and then checks the outcome - if they have heart disease or not. It also calculates the probability for each attribute individually. For example, for a data set with 30 samples if 7 of them had a heart condition, the overall probability is 7/30. Individually, of the seven, 4 had high blood sugar, 4 were male, 6 were above the age of 60, and 5 had severe chest pain. Using both the overall and individual probabilities it is possible to calculate the probability of cardiovascular disease for any combination of the attributes. In this fictitious example, someone who had all four of the same attributes listed has a higher probability of a heart attack than someone who has less. At 86.3% accuracy, this methodology is proven to be the best for predicting heart disease (6).

However, according to the same study, the decision tree algorithm was found to outperform Naive Bayes in regard to predicting those who are not at risk with an accuracy of 89%. The decision tree algorithm is also simplistic and works by creating a tree that follows a path starting with one attribute and splits either when it comes out with a decision or prediction, or when it should be explored with other variables that have an effect. For example, let's say it starts with asking if their blood sugar is above 140mg. If it is, it gives the probability of a heart attack. If not, then it asks if they smoke. For a smoker, it gives the decision, for a non-smoker, it keeps going, and so on (6).

The last algorithm is the KNN, or k nearest neighbors algorithm. It classifies information by the majority of its closest neighbors, so it would find people with similar backgrounds and give them similar diagnoses. Because these algorithms are simple, good at classification, and respond in real time, they play an integral role in data mining and predictions for the health sciences (6).

Health Sciences benefit tremendously from technology. However, due to legal issues and privacy concerns, technological advances are circumscribed. While some argue for the need to protect privacy in the medical industry regarding big data, others argue against it. According to USF health, "Massive amounts of patient data being shared during the data mining process increases patient concerns that their personal information could fall into the wrong hands. However, experts argue that this is a risk worth taking (3)." For example, the coronavirus has captured the news in the past weeks. In order to stop the spread and investigate how and where it spreads, China and Taiwan have resorted to mining and analyzing data, tracking patients and spread patterns. Although encroaching on people's privacy, this approach has managed to control the virus to an extent. By tracking peoples' symptoms and travel history - data they get from QR codes and cell phone movements - the government is able to know who to quarantine and if people are abiding to their self-isolation (7). Although this sounds like spying, the real time tracking of the virus — made possible by the advanced data mining technology available — has helped limit the spread of the coronavirus, saving lives. On the other side, "CEs [Covered Entities, healthcare entities,] remain responsible for a host of

Privacy Rule and Security Rule requirements aimed at safeguarding protected health information. … If a CE fails to comply with an administrative simplification provision, it is directly liable for civil, and in some cases criminal, penalties." (8) Since many legal restrictions are in place protecting the privacy of patient records, in many areas, Healthcare is restricted in its ability to fully optimize the advances made in big data mining.

The use of Big Data Mining in the Healthcare industry has produced more accurate diagnosing, saving lives. However, because of legal and privacy issues, healthcare is still "slow to incorporate the latest research into everyday practice." (9) Even still, predictive diagnosing is becoming more prevalent, leading to questions like, will AI diagnosing take over the role of doctors? Today, this does not appear to be the direction we are heading in, as doctors still need to oversee and correct the diagnoses made by machines. But who knows what tomorrow will bring?

WORKS CITED

(1) "National Ambulatory Medical Care Survey: 2016 National Summary Tables." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, January 19, 2017. https://www.cdc.gov/nchs/fastats/physician-visits.htm.

(2) Prather, J C, D F Lobach, L K Goodwin, J W Hales, M L Hage, and W E Hammond. "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse." Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium. American Medical Informatics Association, 1997. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233405/?page=1.

(3) USF HEALTH. "Data Mining In Healthcare." usfhealthonline.com. USF HEALTH, n.d. https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/.

(4) Kharkovyna, Oleksii. "Artificial Intelligence & Deep Learning for Medical Diagnosis." Medium. Towards Data Science, November 13, 2019. https://towardsdatascience.com/artificial-intelligence-deep-learning-for-medical-diagnosis-9561f7a4e5f.

(5) "Research." Research | Lunit Inc. Accessed March 30, 2020. https://lunit.io/en/research/.

(6) Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction." International Journal of Computer Applications 17, no. 8 (March 31, 2011): 43–48. https://doi.org/10.5120/2237-2860.

(7) Duff-Brown, Beth. "How Taiwan Used Big Data, Transparency and a Central Command to Protect Its People from Coronavirus." FSI. Stanford, March 3, 2020. https://healthpolicy.fsi.stanford.edu/news/how-taiwan-used-big-data-transparency-central-command-protect-its-people-coronavirus.

(8) Pasquale, Frank, and Tara Adams Ragone. "PROTECTING HEALTH PRIVACY IN AN ERA OF BIG DATA PROCESSING AND CLOUD COMPUTING ." STANFORD TECHNOLOGY LAW REVIEW 17, no. 2 (2014): 595–654. https://law.stanford.edu/publications/e-protecting-health-privacy-in-an-era-of-big-data-processing-and-cloud-computing/.

(9) Eliason, Brian, and David Crockett. "What Is Data Mining in Healthcare?" Health Catalyst, July 15, 2019. https://www.healthcatalyst.com/data-mining-in-healthcare.

REVOLUTIONIZING THE FINANCIAL WORLD THROUGH BIG DATA

BY: SARAH WEISS

Finance is an industry that generates huge amounts of data. When collected and analyzed properly, this data has the potential to help financial experts discover patterns, make predictions, and develop strategies that will dictate how billions of dollars move across the global market. Big data is more than just a way to predict share prices and it is completely changing the way the financial world operates.

Big data is the data collected from every social media post, every click on the internet, and every online purchase. With more online interaction in today's world than yesterday's world could ever have dreamed of, there is an endless amount of this data - so much so that it led to the creation of a whole new field of study and technology surrounding how to process it and sort it into useful information. In the finance industry specifically, the emerging field of FinTech, financial technology, focuses on using big data technology to compete with traditional finance methods (1).

One of the areas in finance where big data has had the biggest impact is security and fraud. The finance industry is under constant and significant threat. Location intelligence offers a way to track where a customer uses a product and can provide instant checks that catch unusual or suspicious activity (2). For example, every time a person buys something on a credit card, data is collected about what was purchased, how much was spent, and where the transaction took place. The more often the card is swiped, the more data is collected. If after three weeks of credit card purchases, all made in the same area of Brooklyn, and all within the range of $10-$100, there is suddenly a $1,200 purchase at an Apple store in San Francisco, the system will likely block the purchase from going through. The system learns from the data it collects and can detect any unusual activity. This is a small example, but with petabytes of data being constantly collected, much bigger and more obscure financial threats can be detected and prevented.

Another area of finance, where big data is completely revolutionizing the way things have been done for decades, is the stock market and the world of investment. The market is all about predicting what a company's future will be. This makes it an ideal candidate to benefit from the huge amounts of data being collected on human behavior, social trends, and spending habits. As new technologies emerge that allow for the processing and analysis of big data, it becomes easier to accurately predict a company's earnings before quarterly reports are released. Determining patterns and trends that can influence share prices becomes more feasible (3).

While big data can be exceedingly valuable to the finance world, as with any new innovation, it comes with challenges. One of the key challenges in utilizing big data in finance is that of data quality. Data comes from many different sources, and it's very common to find data that doesn't quite match up, or even contradicting data. It is important to ensure that the data being used to influence important

decisions is accurate and reliable. Another problem with big data that is specific to the finance industry, is the challenge posed by regulatory requirements. The world of finance is laced with strict regulatory requirements, such as the Fundamental Review of the Trading Book (FRTB), that govern access to important data. Therefore, financial data must undergo much more filtering and careful processing than data used in other industries. Finally, there are many ethical questions regarding how the data is collected and how it should be used. New technologies provide the ability to listen in to conversations in private homes, and to monitor web activity. That data can then be used to manipulate the way consumers spend their money, raising many questions on the morality of such data collection. In finance specifically, data can be used to manipulate the markets, affect share prices, and predict earning reports. This brings many ethical issues to the surface (4).

Despite the challenges, big data technology is undoubtedly revolutionizing the entire industry, and will shape the future of the financial world. The world is still discovering its full potential, but with new innovation and tools that have the ability to turn big data into useful information, it is clear that finance will never be the same again.

WORKS CITED

(1) Kh, Ryan. "How Big Data Can Play an Essential Role in Fintech Evolution." SmartData Collective, 24 July 2018, www.smartdatacollective.com/fintech-big-data-play-role-financial-evolution/.

(2) Ewen, James. "How Big Data Is Changing the Finance Industry." Tamoco, Tamoco, 11 Feb. 2020, www.tamoco.com/blog/big-data-finance-industry-analytics/.

(3) Morshadul, Hasan M., Popp József, and Oláh Judit. "Current Landscape and Influence of Big Data on Finance." Journal of Big Data, vol. 7, no. 1, 2020. ProQuest, https://search.proquest.com/docview/2376143296?accountid=14375, doi:http://dx.doi.org/10.1186/s40537-020-00291-z.

(4) Pearlman, Shana. "Big Data in Finance: What, Why, and How - Talend." Talend Real-Time Open Source Data Integration Software, 8 Aug. 2019, www.talend.com/resources/big-data-finance/.

MACHINE LEARNING
MAKING BIG DATA USEFUL

BY: RINA YOUNGER

Humans are great at many things. We are great at ordering pizza. We are great at losing our keys. We are great at sleeping through our alarm clocks in the morning. But one of the things that we are the greatest at is generating data. Every time you search the weather on google, order something on Amazon, or post a selfie on Instagram, you generate data. In fact, humanity as a whole produces 2.5 quintillion bytes of data every day (1). That's a lot of data!

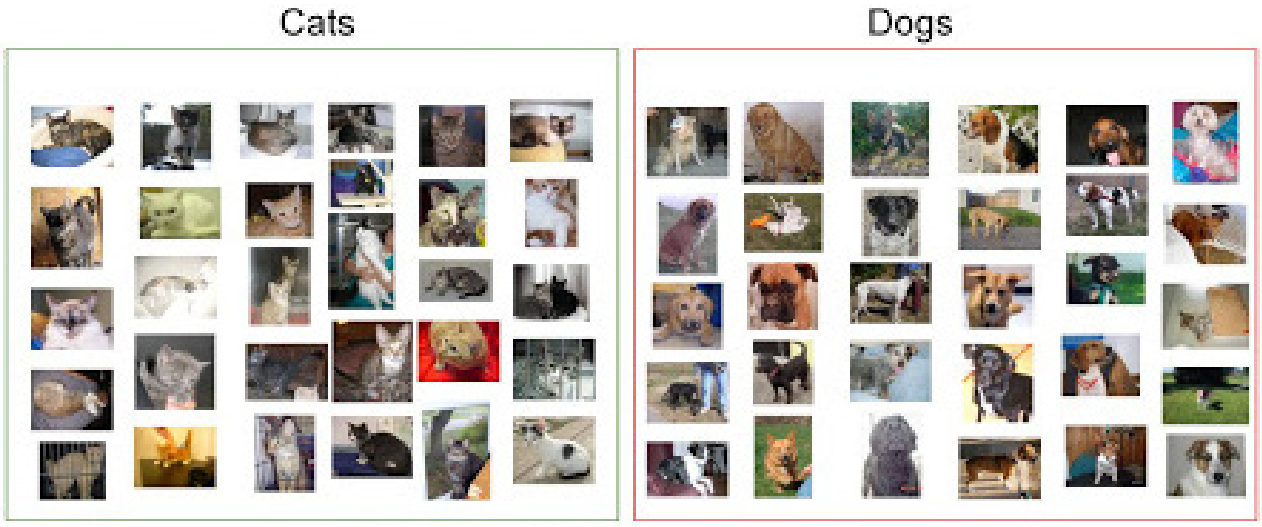
Much can be learned from analyzing the data that we produce. Machine learning is a technology that offers a way to find meaning in big data (2).

But what is machine learning exactly? "Machine learning is the science of getting computers to act without being explicitly programmed" (3). The implications of this definition are baffling! How can you get a computer to do something you didn't tell it to do? Computers cannot think. If a programmer does not explicitly program a computer to do something, then the computer will not do it. A computer cannot decide on its own to do something. Or can it? The definition of Artificial intelligence is "the capacity of a computer to perform operations analogous to learning and decision making in humans" (4). Machine Learning, a subset of artificial intelligence, is a technology that allows computers to learn from data and make decisions by themselves (5).

Traditionally, a computer is provided with data and rules. Data is the input the computer will receive, and the rules are written by programmers. Programmers write code that instructs the computer as to how to react to different data sets provided. The output that the computer produces depends on the input provided and the rules in place.

Machine learning does just the opposite. Machine learning does not follow the traditional way of providing the computer with data and guidelines and expecting the computer to produce the correct answer. Instead, Data and the expected output (commonly known as labels) are provided to the computer, and it's the computers job to infer what the rules are (6).

A classic example of a machine learning task would be getting a computer to differentiate between pictures of cats and dogs. The traditional approach to solving this problem would be to give the computer extensive guidelines for how to differentiate between cats and dogs. For example, cats generally have pointier ears than dogs, cats have longer whiskers than dogs, and dogs have snouts, etc. A programmer would have to compile a thorough list of rules for every scenario to enable a computer to discriminate between pictures of cats and pictures of dogs. Machine learning circumvents the hassle of gathering all these rules. Instead of providing the computer with the rules for how to categorize these pictures, we provide the computer with labels to what animals the pictures are of. Then we give the computer the task of finding the patterns and connections that are common to all the pictures of dogs versus those of the cats (7).

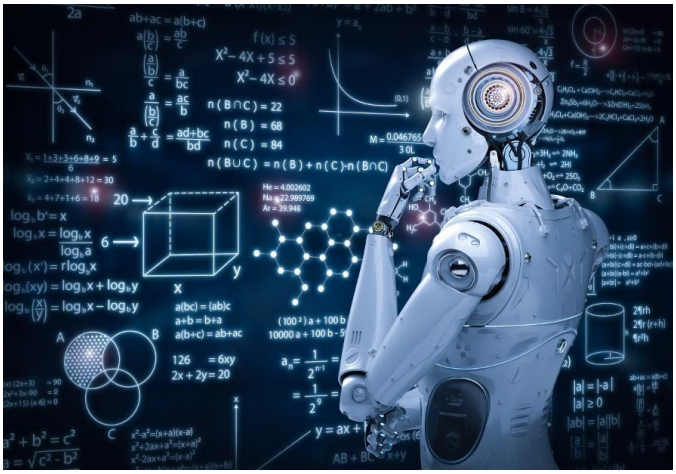


By going through large data sets of pictures of cats and dogs and seeing how each one is characterized, the computer slowly learns which features and patterns make up a cat and which make up a dog. After this training period the computer is now capable of working out whether a picture from a new data set is either a cat or a dog.

This technique describes a method of data analysis called supervised learning. Supervised learning is when data and the correct label of the data are provided to the computer. In our example, the pictures of the animals and the correct classification of the animals were both provided to the computer. But there is another, even cooler, machine learning technique. Suppose we were really lazy. Not just too lazy to give the computer a list of guidelines to follow, but so lazy we don't even give the computer the labels for the data we will provide. We will just give the computer a large data set and let the computer come up with patterns and predictions on its own (8). This is called unsupervised learning. Unsupervised learning is when the computer is only given a data set, without any labels or rules. The computer is tasked with coming up with a way to group the data all by itself. Surprisingly, this works! Machine learning algorithms are so sophisticated that they can find patterns and connections within data that people didn't even know existed.

This pretty much sums up the core of machine learning: you get a set of data with inherent patterns, and you have the computer figure out what those patterns are (9). Deciphering pictures of cats and dogs may not seem that impressive to human beings with the innate capability to recognize different animals. But classifying cats and dogs is just a simple example of what machine learning can do. Computers are much faster than humans and can analyze massive amounts of data. The ability to find patterns among large data sets gives computers the capacity to surpass humans in their understanding of data.

Machine learning is a powerful tool that helps engineers analyze large data sets, and make predictions based on what is learned. Self-driving cars, accurately predicting whether a skin mole is cancerous, and being able to recognize and understand human speech are all made possible through machine learning (10). The amount of data we have at our disposal is immense, but the data we have collected is only as useful as the analysis tools we have to understand it. Machine learning offers a way to analyze and learn from big data in a meaningful way. The data that we generate every day does not have to sit idly and wastefully in a database somewhere. With machine learning, big data can be evaluated and utilized in a way that makes our world smarter, safer, and happier.



WORKS CITED

(1) Marr, Bernard. “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.” Forbes, Forbes Magazine, 5 Sept. 2019, www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#310ddbfe60ba.

(2) Google Cloud Platform. “What Is Machine Learning? (AI Adventures).” YouTube. August 24, 2017. https://www.youtube.com/watch?v=HcqpanDadyQ

(3) “What Is Machine Learning? - Introduction.” Coursera, www.coursera.org/lecture/machine-learning/what-is-machine-learning-Ujm7v.

(4) Artificial Intelligence.” Dictionary.com, Dictionary.com, www.dictionary.com/browse/artificial-intelligence.

(5) Ronald, Jasmine. “Machine Learning vs Big Data: Let's Find the Relationship between Them.” Medium, Towards Data Science, 15 Oct. 2019, towardsdatascience.com/machine-learning-vs-big-data-lets-find-the-relationship-between-them-e55c9c861311.

(6) TensorFlow. “Intro To Machine Learning (ML Zero To Hero, part 1).” YouTube. August 30, 2019. https://www.youtube.com/watch?v=KNAWp2S3w94

(7) Olckers, Ayran. “How To Teach a Computer to Distinguish Cats From Dogs.” Medium, Medium, 16 May 2017, medium.com/@TheGeekiestOne/how-to-teach-a-computer-to-distinguish-cats-from-dogs-d66cc0679287

(8) Richard BradleyFollowDigital Core Transformation | Consumer & Industrial Products | I4.0 Value CreationLike30Comment6ShareLinkedInFacebookTwitter2, and Follow. “Artificial Intelligence Explained.... with Cats and Dogs...” LinkedIn, www.linkedin.com/pulse/artificial-intelligence-explained-cats-dogs-richard-bradley/.

(9) TensorFlow. “Intro To Machine Learning (ML Zero To Hero, part 1).” YouTube. August 30, 2019. https://www.youtube.com/watch?v=KNAWp2S3w94

(10) TED. "What AI is -- and isn't | Sebastion Thrun and Chris Anderson." YouTube. December 21, 2017 https://www.youtube.com/watch?time_continue=627&v=J6tgYBMXR6s&feature=emb_logo

DON'T MISS THE BOAT

EXPLORE THE OPPORTUNITIES OF NATURAL LANGUAGE PROCESSING

BY: CHANA FINE

A sea of information has accumulated from the constant interactions between humans and technology. Big data refers to the large amount of structured and unstructured data that overwhelms businesses on a daily basis. Big data can be used to gain insights that lead to better decisions and developments. With the growth of the Internet of Things, data is collected from a variety of sources. Artificial intelligence relies on big datasets that reflect patterns and contribute to the search for answers to the biggest questions in our generation. The insights captured from big data drive innovation (1).

Natural Language Processing (NLP) is a type of artificial intelligence that teaches computers how to understand, interpret, and manipulate human language. NLP stems from computer science and computational linguistics. The goal of NLP is to bridge the gap between human communication and computer understanding (2).

The science of natural language processing has rapidly advanced with the increased interest in human-to-machine communications, availability of big data, powerful computing, and enhanced algorithms. Humans interact with each other in native languages like English, Spanish, or Chinese. Computers don't speak the same language. They are designed to understand machine code, which looks like a bunch of zeros and ones. This language is a far outcry from natural languages that people are accustomed to (2).

In order to make computers versatile to human needs, communications between humans and computers is critical. In the 1950s, early computers did not understand any languages besides machine language. Programmers used punch cards with certain patterns of punches to symbolize the syntax of machine language. Today, many platforms support human voice control to customize the digital experience (2).

Natural language processing has two branches: natural language generation and natural language understanding. Natural language generation (NLG) systems convert information from databases into readable human language. Natural language understanding (NLU) systems convert human language into representations that computer programs can easily manipulate (3). NLG writes natural language and NLU understands it (4).

Natural Language Generation (NLG) is the process of generating meaningful phrases and sentences in natural language form. It automatically produces narratives that are based on structured data. NLG makes raw data understandable to humans. Writing text like financial reports, product descriptions, and meeting memos becomes incredibly easy. NLG alleviates the burden from analysts to summarize data and write reports tailored to the audience. NLG can be used in client-facing applications in administrative platforms including analysis for business intelligence dashboards, reports on business data, reports on IoT device status and maintenance,

individual client portfolio summaries and updates, and personalized customer communications (4).

Google Assistant uses NLG to answer questions about yesterday's football game, or summarize the contents of an email. Research is being conducted to use NLG to generate news articles, especially in finance (5). When creating an NLG system, developers outline the format of the content. Every content type has a unique structure. Social media posts are short and concise. Poems may have short lines and flowery language. Financial reports contain technical jargon and numeric data. Structured data is fed into the software and processed with conditional logic. NLG systems can process large datasets and produce narratives a thousand times faster than humans. The Associated Press uses NLG to turn the sports data that interests readers into news articles that are written in the appropriate lingo (6). Another widely used application of NLG is chatbots (5).

Natural language understanding (NLU) is artificial intelligence focused on recognizing patterns and meaning within human language. With NLU, computers can deduce the meaning behind the speaker's words. This voice technology enables Alexa to infer that a weather forecast is being requested when it hears "Alexa, what's it like outside?" The phrase does not contain the word "weather" yet Alexa has been programmed to understand the meaning behind the literal message. Speaking to a computer feels like an actual conversation when the human does not need to follow a specific structure. NLU enables computers to extrapolate the speaker's intention regardless of how it was phrased (7).

NLU provides computers with the required context for human speech and the flexibility to understand many variations of identical messages. Before NLU, speech-enabled weather apps required the user to ask questions in a predictable structure, such as "is it raining." When users can ask for the weather with different expressions, their experience is more intuitive and delightful. Voice technologies that incorporate NLU can be used to create more productive user experiences (7).

With the capabilities to store big data, intelligent algorithms can be applied to the datasets to produce natural language processing that enhance user experiences. Interactions between humans and computers can become seamless digital experiences thanks to natural language generation and understanding. With the awareness of the plethora of tools that utilize artificial intelligence, challenge yourself to spot computer-generated text. As the "natural" capabilities of natural language processing continue to improve, this challenge may become increasingly more difficult. For all you know, this article may have been written by a high-powered algorithm. While you continue to ponder, I will enjoy personalized cruise recommendations from my phone's voice command. Bon Voyage!

WORKS CITED

(1) "Big Data: What It Is and Why It Matters." SAS, www.sas.com/en_us/insights/big-data/what-is-big-data.html.

(2) "What Is Natural Language Processing?" SAS, www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html.

(3) "Big Data & Natural Language Processing", Big Data News, https://www.bigdatanews.datasciencecentral.com/profiles/blogs/big-data-natural-language-processing

(4) "A Comprehensive Guide to Natural Language Generation", Sciforce, https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548

(5) "Google talks NLG and what it can do for businesses", Tech HQ, https://techhq.com/2018/05/google-talks-nlg-and-what-it-can-do-for-businesses/

(6) "Artificial Intelligence can Now Write Amazing Content -- What Does That Mean For Humans?", Forbes https://www.forbes.com/sites/bernardmarr/2019/03/29/artificial-intelligence-can-now-write-amazing-content-what-does-that-mean-for-humans/#1cceae5d50ab

(7) "What is Natural Language Understanding (NLU)?" Amazon, https://developer.amazon.com/en-US/alexa/alexa-skills-kit/nlu

BIG DATA IN THE CLOUD

BY: KARA NOBLE

The term cloud computing was coined back in 2006 when big companies like Google and Amazon started accessing resources like software, file storage and computer power over the Web instead of on their local machines (1).

The "cloud" is a fanciful term used to describe powerful servers accessed over the internet, which manage databases and software. Cloud computing helps companies store data effectively without having to manage the necessary hardware on their own. The cloud is also useful because the same data can be accessed by any machine, granting portability and increasing availability, enabling businesses to access their data from anywhere (2). Cloud computing is what enables a user to log in to their email from any machine and not only send a new email but review old emails and current drafts as well.

Big data is about collecting and analyzing data on a massive scale. Cloud Computing is the infrastructure behind managing software and data. Cloud computing and big data make a great pair in three ways: improved analysis, simplifies infrastructure and lower costs (3).

Big data analysis is improved through using cloud computing. Using the cloud enables data to be integrated from various sources and be analyzed as one unit, improving analysis and getting better results (3).

Cloud computing simplifies the necessary infrastructure for big data. The cloud provides a robust infrastructure that can handle the vast amount of data that filters through a system, unlike most traditional infrastructures that can't keep up with the volume and pace. Cloud computing is also flexible and can scale to the needs of each job, managing all kinds of workloads (3).

Businesses incur lower costs and expenses by outsourcing to cloud vendors. Cloud computing offers large-scale services to small companies that otherwise wouldn't have been able to provide their own infrastructure to handle their needs (3).

Let's dive a little deeper into the 5 V's of Big Data and how cloud computing can be a useful resource.

Volume – Big data is all about managing enormous amounts of data. Since the cloud provides limitless storage (assuming you are willing to pay for it), it becomes an appealing option to businesses with growing data needs. Moving data and analytics provides users with flexibility and scalability. It is important to realize that costs can add up and users should be careful not to store unnecessary data in the cloud (4).

Variety – in big data terms, variety refers to heterogeneous sources. For example, different departments of a company may want to analyze different types of data – pictures versus reports, etc. The

cloud accommodates these different analyses (4).

Velocity is the speed that data is collected, which can become complicated as companies try to provide the enormous amount of power necessary to handle the continuous and huge amount of data. The cloud can scale to meet the needs as the system grows (4).

Veracity refers to the inconsistencies and uncertainties in the data. The cloud provides more room for the user to make a big mess and to compromise quality data (4). This area should be carefully reviewed before a company makes the move to cloud analytics.

Value is the edge that the business acquires as a result of this data collection. The cloud provides means to analyze the data and turn it into useful information that a company can use to make more informed decisions and take better steps to improve their services (4).

While most can agree that cloud computing and big data are an excellent pair, each business needs to determine for itself if switching to cloud computing will raise the bottom line and add enough value to the company to be worth the cost.

WORKS CITED

(1) Regalado, A., 2020. Who Coined 'Cloud Computing'?. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing/> [Accessed 15 March 2020].

(2) 2020. [online] Available at: <https://www.cloudflare.com/learning/cloud/what-is-the-cloud/> [Accessed 15 March 2020].

(3) Whizlabs Blog. 2020. Big Data And Cloud Computing – A Perfect Combination - Whizlabs Blog. [online] Available at: <https://www.whizlabs.com/blog/big-data-and-cloud-computing/> [Accessed 15 March 2020].

(4) SearchCloudComputing. 2020. Cloud Cost Implications Of The 5 V's Of Big Data. [online] Available at: <https://searchcloudcomputing.techtarget.com/tip/Cost-implications-of-the-5-Vs-of-big-data> [Accessed 15 March 2020].

BIG DATA AND HADOOP

BY: AARON FARNTROG

Before addressing the elephant in the room, let us discuss big data, a term used to describe the collection of huge amounts of data that can grow exponentially over time. The characteristics of big data include volume, variety, velocity and veracity, the four V's.

Volume refers to the incredible quantity of data that can be generated each second from single or multiple sources, such as cell phones, social media, online transactions, etc.

Variety refers to the different types of data such as structured, semi-structured, and unstructured. Structured data has a fixed format and size, semi-structured data has a structure but cannot be stored in a traditional database, and unstructured data does not have any specific format, which hinders analysis.

Velocity refers to the speed at which data is generated and processed. Big data is often available in real-time. Compared to small data, big data are produced more continually.

Veracity refers to the inconsistencies within the data, thus hampering the process of handling and managing the data effectively.

An example of big data would be the collection of emails. According to Statista, there will be an estimated 126 trillion, that's trillion with a 't', emails sent in 2020. How's that for volume? Emails contain a tremendous variety of data, which can include images, pdfs, text, and videos. Additionally, emails have tremendous velocity. There are an estimated 306 billion emails sent daily, which averages to 3.5 million emails per second.

Due to the four V's of big data, storing, processing and analyzing data using traditional systems is extremely difficult. Serial memory cannot be used to process large input streams. In addition, a single central storage location cannot store enormous volumes of data. Lastly, traditional systems are inadequate and do not have the capability to process unstructured data. In conclusion, a traditional relational database management system does not have the capacity to store massive, complex data, and single processes do not work efficiently with big data.

To solve these challenges, Doug Cutting and Mike Cafarella created Hadoop the Yellow Elephant. Hadoop is a collection of open-source programs and procedures which can be the backbone of any big data operation. The Hadoop system runs on commodity, affordable, servers that store and compute massive amounts of data. The Hadoop framework is made up of three core components: Hadoop Distributed File System (HDFS), MapReduce, and Yet Another Resource Locator (YARN).

Hadoop Distributed File System (HDFS) provides a distribution method that stores the data in blocks with a user-specified size, which are themselves stored inside DataNodes. Suppose you have

512 MB of data and you have configured HDFS such that it will create data blocks of 128 MB. HDFS will divide the data into four blocks, 512/128=4, and store it across different DataNodes. Additionally, the data blocks are replicated on different DataNodes to provide fault tolerance.

MapReduce is a software framework that helps in writing applications that process large data sets using distributed and parallel algorithms inside HDFS. A Map function filters or sorts data and the Reduce function performs a summary, or grouping, on that data. The big advantage is that when you run your program through MapReduce, it is automatically run on all nodes in a Hadoop cluster so you can process data in a divide and conquer approach.

Yet Another Resource Locator (YARN) is the resource management and job scheduling technology in the Hadoop distributed processing framework. It is responsible for allocating system resources and scheduling tasks within the Hadoop ecosystem. In shared environments like Hadoop, you don't want any program to hog resources and overwhelm the system, so YARN is used as a way to play nice with the other programs. Essentially YARN is the brain of the Hadoop Ecosystem.

With Hadoop, all the problems of processing and collecting colossal amounts of data are solved. HDFS is highly scalable because data is distributed over several machines, and any number of machines can be added at any point. We no longer have the bottlenecks of a traditional RDBMS. Additionally, since data in HDFS is stored as files, Hadoop does not require a schema or a structure for the data that will be stored. Therefore, Hadoop can be used to store any unstructured data - thereby solving our second problem by allowing the storage of heterogeneous data. Lastly, MapReduce runs its processes on all nodes utilizing the power of parallel processing. By switching from traditional serial processing to parallel processing, we can cut down the processing time significantly allowing for efficient handling of big data.

In today's competitive environment, companies in all industries harness the power of big data to create useful insights and gain a competitive advantage from the increasing flood of generated data. With emerging big data industry trends in almost all sectors, there is an increasing demand for Hadoop developers. With a little background in Java, it would be a good idea to take a deep dive into learning the Hadoop ecosystem to show employers your understanding of how big data is managed.

WORKS CITED

(1) Clement, J. "Daily Number of e-Mails Worldwide 2023." Statista, 9 Aug. 2019, www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.

(2) "Apache Hadoop." Wikipedia, Wikimedia Foundation, 26 Mar. 2020, en.wikipedia.org/wiki/Apache_Hadoop#History

(3) "Top 10 Industries Using Big Data and 121 Companies Who Hire Hadoop Developers." DeZyre, www.dezyre.com/article/top-10-industries-using-big-data-and-121-companies-who-hire-hadoop-developers/69.

SEARCH ENGINES

BY: CHAYA SARA ZITWER

What is a Search Engine? The Web can be viewed as a single large database, or a group of smaller databases. However, the Web is an ineffectively designed database that has no simple or consistent method of query, such as the Structured Query Language (SQL) used in many relational databases. This is often because the Internet database is not only a database, but also an information base. Internet searches need to be organized not only by the address of data, but also by the meaning of the documents. Due to the lack of consistent design in the "Internet database", many search engines use different techniques to query the same information base, sometimes giving very different results (1).

A Search Engine is a web-based tool that allows users to locate information on the Web. A search engine can have many different components, such as the spider, the index, and the algorithm.

Web Spiders, Bots, and Crawlers

Search engines use robots to construct their indexes. Robots are applications that recurrently search the web and search for modifications to update the database. Starting at a particular web page, the robots make a list of all the links on that page. Those linked pages are then handled to find links to other pages, and so on. Finally, the robot has links to the bulk of the content on the Web. Some content, though, is tougher to locate. Some pages can't be reached from a website's home page, but are instead found through the site's own search engine. This unlinked content is known as the deep web. Incorporating deep web content into a search engine index typically requires some help from the site. Site managers have numerous approaches to provide web-crawling robots with a "table of contents" for all the pages on the site. One such way is a document referred to as a Sitemap. This document is named after the site map page some sites provide for users to quickly discover the content they are looking for, but has a specific format that's easy for robots to process. Sitemaps keep search engines updated with content changes and are especially useful for sites with deep content that would otherwise be left out of search engine indexes (2).

The Index

The index, sometimes called the catalog, stores a copy of every web page with a document profile. The process of indexing the documents involves taking the documents located by spiders and translating them into the information retrieval language (IRL). The result of this translation is referred to as a document profile. The document profile is used for comparison with the search query formulation and the selection for the output. If there is any change to a document, the contents of this catalog are updated with new information as well. A document does not become available to the search engine until it has been indexed (1).

The Algorithm

Every search engine uses different formulas to produce search results. The algorithm is a complex equation that calculates a value for any given site in

relation to a search term. The results for the query are then displayed on the SERP (search engine results pages). Every search engine's algorithm is unique, so a top ranking on one search engine does not guarantee a prominent ranking on another search engine. The algorithms used by search engines are closely guarded secrets and are continuously changing. Therefore, the criteria to optimize a site must be conjectured through observation, as well as trial and error. Search engines use the basic HTML structure to determine relevance. Large photos, or dynamic Flash animation mean nothing to search engines, only the actual text on the pages do (3).

Search Engine Optimization

Search Engine Optimization (SEO) is the task of making web pages rank higher in search engines. Some optimizations can include technical SEO, On-Site SEO, and Off-Site SEO. Technical SEO involves changing the site's settings to ease the job of search engine crawlers. On-Site SEO includes the structure of the web page, for example, using the right html, css, etc. to send the right indicators to search engines. And lastly, Off-Site SEO involves the number of incoming links, also known as backlinks, for the website. Whether a site is reliable or not is decided by the number and quality of backlinks it receives from other websites. By having good content, a website will get natural backlinks, which can make it rank higher (4).

Google

Google is a search engine that was started in 1996 by Sergey Brin and Larry Page as a research project at Stanford University to find files on the Internet. What helps Google stand out from its competition and be the top search engine is its PageRank technique. PageRank sorts web pages by applying them with an algorithm and giving it a numerical weight. This "weight" is then applied to sort results when Google's search engine is used (5).

According to Internet Live Stats, "Statistics say that Google now processes over 40,000 search queries every second on average, which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide" (6).

Big data has altered the world, and SEO is one of the most influenced by the big data revolution. Big data has benefitted SEO by giving search engines what to analyze and by making it easier to implement search engine optimization techniques (7).

WORKS CITED

(1) Gong, Juan. Internet Search Engine Technologies and Search Optimization, Fordham University, Ann Arbor, 1999. ProQuest, https://search.proquest.com/docview/2203348853?accountid=14375.

(2) V., Anton Spraul. How Software Works : The Magic Behind Encryption, CGI, Search Engines, and Other Everyday Technologies, No Starch Press, Incorporated, 2015. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/touromain-ebooks/detail.action?docID=4503164.Created from touromain-ebooks on 2020-03-15 07:54:31.

(3) DesignHammer Website Design and Development in North Carolina. 2020. What Is A Search Engine?. [online] Available at: <https://designhammer.com/services/seo-guide/search-engines> [Accessed 15 March 2020].

(4) Chris, Alex, et al. "What Is SEO (Search Engine Optimization) And Why Is It Important." Reliablesoft.net, 6 Nov. 2019, www.reliablesoft.net/what-is-search-engine-optimization-and-why-is-it-important/.

(5) " What Is Google?" Computer Hope, 6 Mar. 2020, www.computerhope.com/jargon/g/google.htm.

(6) "Google Search Statistics - Internet Live Stats." Internet Live Stats, www.internetlivestats.com/google-search-statistics. Accessed 19 Mar. 2020.

(7) The Link between Good SEO and Big Data. (n.d.). Retrieved from https://datafloq.com/read/link-between-good-seo-big-data/4479

DATA ANALYSTS

BY: PENINA ZIEGLER

We live in a world that is focused on statistics. The job of a data analyst is to inspect large amounts of data in order to help companies make decisions. The analysis process includes defining the issue, organizing the data, and reporting the results. The job of data analysts can be applied to many fields.

Data can be analyzed for medical purposes . Questions such as "Is being overweight really related to future mortality?" are explored by data analysts in the medical field. The NCHS has been surveying the health of the public since the 1960's . They look at the responses to surveys in previous years and make comparisons to the National Death Index to track what happened to those surveyed. As an example, data analysts have checked to see if those surveyed in 1987 who responded that they are smokers were likely to die by 2007 due to smoking. These analysts provide reports and knowledge that help shape the medical research fields. Using the statistics they have found, analysts warn the public of the dangers surrounding unhealthy habits such as smoking (1).

Utility companies are another sector that has data analysts working for them. I had the opportunity to interview Sherry Blassberger, a data analyst working for a utility company. She described the important job functions of the data analysts, which are spread out across the company. These analysts provide analysis in areas such as electric construction, human resources, gas operations, etc. She works as a data analyst for electric operations and reviews crew start times and end times and other factors such as job delays to produce reports that help drive crew efficiency. For example, if they find that construction crews are often delayed waiting for a flush crew to flush out the manhole, then process changes are made so that flushes are called ahead of the construction crews to minimize delays.

Sherry Blassberger went on to describe other critical data analysis that she has performed that led to changes in the business. The supervisors approve time for the crews in the work management system. To ensure that the hours are correctly reported and employees do not lose pay, Sherry crunches the data of over 1,000 employees to determine inaccuracies in time reported and drive action plans for recurring issues to prevent future losses in pay.

Yet another area where data analysts are heavily relied upon is that of the stock market. A specialty position is the Quantitative Data Analyst otherwise known as "Quants." These data analysts work alongside stock traders and develop complex algorithms to analyze data on expected returns of stocks. Data analysts weigh risks and include historical data and current trending data to report on the risks of stock portfolios. The data gathered by the Quantitative Data Analyst drives decision making for the stock traders and is considered an invaluable position in the world of dynamic daily stock trading (2).

Accounting departments often have accountants work alongside data analysts. Companies are

under pressure to increase revenue and decrease costs. The analysis performed by data analysts provides critical information to improve the company and ensure its success. The McKinsey & Company 2013 Datamatics survey shows the need for data mining and analysis. The survey estimated that companies that perform extensive data mining and Analysis are 2.3 times more likely to have above-average sales, 2.3 times more likely to generate higher returns on investments, and 21 times more likely to have increased profitability. This is because data analysts can use data mining on customer cost and profit data to uncover patterns and figure out how to increase profitability which then provides a tremendous edge to the company's success (3).

Data analysts benefit from strong mathematical skills. Actuaries can be viewed as data analysts since they perform analysis on data and provide risk assessments. The actuarial profession exists in many areas such as property risk, auto insurance risk, flood risk, etc. Insurance guarantees that a premium paid will cover a future event. Actuaries must analyze historical data to make sure that enough money is charged for the rare future events that may occur. Actuaries use the data analysis to determine premiums charged by insurances (4).

Data analysts can be very valuable to a company. Whether it is a medical, utility, stock trading, accounting, or insurance company, data analysts can help provide information for the benefit of the company. Therefore, data analysis is becoming a popular field that is continuously growing.

WORKS CITED

(1) Vangelova, Luba. "Data Analyst." Science Teacher, vol. 82, no. 5, Summer 2015, pp. 60-61. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&AuthType=ip,sso&db=eft&AN=103336132&site=ehost-live.

(2) Kakushadze, Zura, and Willie Yu. "Decoding Stock Market with Quant Alphas." Journal of Asset Management, vol. 19, no. 1, 2018, pp. 38-48. ProQuest, https://search.proquest.com/docview/1994551518?accountid=14375, doi:http://dx.doi.org/10.1057/s41260-017-0059-2.

(3) Pickard, Matthew D., and Gary Cokins. "From Bean Counters to Bean Growers: Accountants as Data Analysts -- A Customer Profitability Example." Journal of Information Systems, vol. 29, no. 3, Fall 2015, pp. 151-164. EBSCOhost, doi:10.2308/isys-51180.

(4) Owen, Rebecca. "Actuaries Are Paying Attention to Climate Data." Bulletin of the American Meteorological Society, vol. 100, no. 1, Jan. 2019, pp. S5-S8. EBSCOhost, doi:10.1175/BAMS-D-18-0293.1.

BIG DATA CHALLENGES

BY: TEHILA RAFUL

Big data can be the solution to the problems of many large corporations today. Big data allows companies to store and analyze large amounts of data. With today's large companies and organizations, big data is very helpful.

While big data offers many advantages for large companies, it still poses many challenges. Many companies that have invested in big data are now facing different challenges. In fact, the IDG Enterprise 2016 Data & Analytics Research found that 90 percent of those surveyed, reported running into challenges related to their big data projects (1).

So, what caused 90 percent of people to report that they ran into challenges when dealing with big data?

The most common challenge of big data is knowing how to store such large quantities of data efficiently. Companies that deal with social media, email correspondences, and economics all have lots and lots of data that need to be stored. The storage place must also have room for data growth as data amounts increase. IDC estimates that the amount of information stored in the world's IT systems is doubling about every two years (1).

The data also needs to be processed efficiently. For example, a company like Facebook is constantly receiving information from its customers. Facebook needs to be able to store all this information as well as process it so it can extract the valuable information and then use that information.

Another challenge of big data is processing the information quickly and in a timely manner. Companies that receive a constant flow of information need to be able to update their databases every second. As PwC's Global Data and Analytics Survey 2016 found, "Everyone wants decision-making to be faster, especially in banking, insurance, and healthcare (1)." This can be very challenging for big data because there is so much information that needs to be processed.

Another factor that poses a challenge to big data is the variety of data formats. Data formats vary from images and videos to emails and word process documents. All these different formats need to be processed accordingly. This makes it harder to process the information quickly since it is hard to compare information between an image and an email (1).

Additionally, certain data structures must be built to hold all the different formats of information. While this may not be a challenge specifically for big data it is still something to take into consideration (2).

Another aspect of big data that may be challenging is securing all the data. With so much information concentrated in one area, it can be an enticing target for hackers. In addition, since there is so much information that needs to be processed, the data is processed through different venues. This can lead to

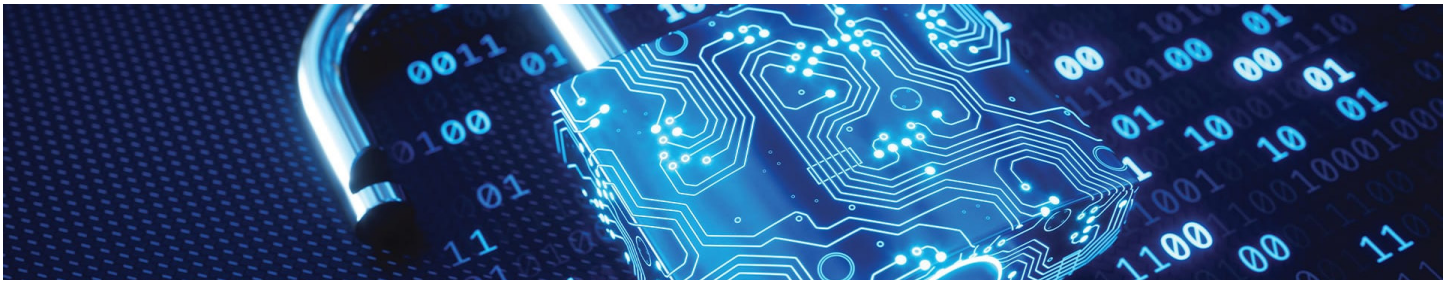
leaking information and data that is not so secure. Certain security measures should be taken when dealing with big data (2).

To conclude, big data is what today's large corporations are using to store and process their data. While big data offers many advantages, the challenges also need to be considered.

WORKS CITED

(1) Harvey, Cynthia. 2017. "Big Data Challenges." datamation.com. June 5. https://www.datamation.com/big-data/big-data-challenges.html

(2) Vaghela, Yuvrajsinh. 2018. "Four Common Big Data Challenges." dataversity.net. June 28. https://www.dataversity.net/four-common-big-data-challenges/.



BIG DATA'S ROLE IN NATIONAL SECURITY

BY: HANAH LAVIAN

Technology is used everywhere, from PCs, phones, and vacuum cleaners, to nuclear bombs. The technology boom in the past 30 years has led to a new form of criminal activity. Cybercriminals are a threat to national security but can be stopped with the help of big data. The abundance of devices and websites that are used helps Homeland Security and Public Safety collect and analyze enormous amounts of information that they can use to their benefit.

Most online activities and interactions are not important to the National Security Agency (NSA). They have to use analytic tools to sift through the vast amount of big data they have access to, because most of the data the NSA receives is irrelevant to them. Analyzing big data in many ways such as: anomaly detection, association, classification and clustering, link analysis, and machine learning, help maintain the nation's security (1).

Anomaly detection is when a data analyst notices an unexpected change in the normal pattern in the data he/she is analyzing. This ensures that any suspicious online activity, even by internal employees, is monitored and reported. Association is another way that big data is used to ensure National Security. Association mining algorithms discover interesting relationships and patterns hidden in big data. Association detects suspicious groups of people, organizations, and locations which can then be monitored (1).

Classification and clustering is another common tool used to analyze big data to ensure national security. Classification algorithms group data from previous instances to detect future cases that are of similar categories or classes. Classification collections can be used to distinguish an innocent intercepted phone call from a suspicious interception. Clustering is an analysis tool that groups similar data together. Then, big data is organized according to topics so that they can be found easily. This helps the NSA sift through the abundance of big data collected from sources, such as social media, more effectively (1).

Ex-criminal hacker Jon Miller ensures criminals, like his former self, cannot attack companies and government agencies. He is able to track criminal and terrorist activity and all those associated with them by using classification and clustering techniques (2). Data clusters are built using a criminal's history of posts on social media. Those data sets can be used to compare other social media accounts to detect future criminals. According to Miller, "Ideally, when somebody goes out and does something like a shooting, you should be able to plug that user's social media presence into a machine to say, 'These types of posts can lead to this type of event' ... and then the machine would perform large-scale psychological profiling on the public" (2).

Furthermore, link analysis is a technique used to ensure national security. Link analysis algorithms detect relationships between suspicious activities, which helps identify networks of terrorist organiza-

tions. Link analysis plays a major role in discovering terrorist or criminal networks, such as al-Qaeda through social network analysis (1).

Machine learning is another useful tool to ensure national security. Machine learning algorithms can, for instance, retrieve hidden context from document collections, identify phishing attacks, detect network intrusion, recognize human faces and analyze crowds. Machine learning can automate security through creating a set of algorithms that can be changed according to the big data they process. However, the results found through analyzing big data through applications and algorithms cannot draw conclusions on its own. Human judgment is necessary to identify the context of the analysis. Big data cannot replace humans' roles in national security, such as thinking, asking questions, and making judgements (1).

Cyberspace is a growing threat to American security. According to the Department of Homeland Security, "Sophisticated cyber actors and nation-states exploit vulnerabilities to steal information and money and are developing capabilities to disrupt, destroy, or threaten the delivery of essential services" (3). Today, nations can go to war with each other exclusively through cyberwarfare. Cyber-attacks have led to an expense of $2.1 trillion to the global economy in 2019, more than four times the cost in 2015 (4). The U.S. government tries to combat cyberwarfare and keep the country secure in many ways. Cybercriminals are especially difficult to arrest; the United States can only make arrests in another country if the country has an extradition treaty with the U.S. Otherwise, the U.S. must wait for the criminals to travel out of the country (5).

While America is trying to hack the databases of enemy countries, enemy countries are doing the same. Unlike regular warfare, a cyberattack can be initialized by anyone. President Trump says, "It could be Russia, but it could also be China. It could also be lots of other people. It also could be somebody sitting on their bed that weighs 400 pounds" (6).

Cybercriminals and terrorists present a major threat to our national security. With the correct tools and algorithms, big data can be used to ensure our safety in the US.

WORKS CITED

(1) Puyvelde, Damien Van, et al. "National Security Relies More and More on Big Data. Here's Why." The Washington Post. 27 Sept. 2017. 15 Mar. 2020.

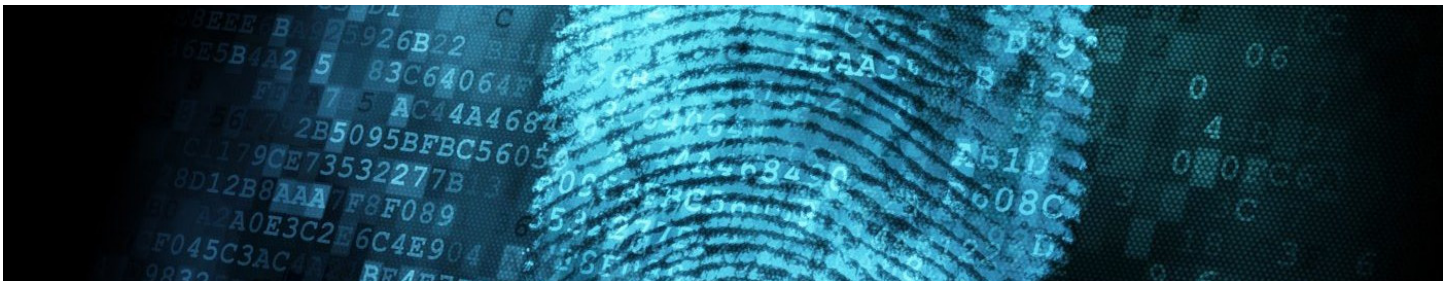
(2) Vasan, Paula. "Secretive Firm That Tracked Bin Laden on a New Manhunt." CNBC. 16 Oct. 2015. Web. 14 Mar. 2020.

(3) "Cybersecurity." Homeland Security, https://www.dhs.gov/topic/cybersecurity, Web. 25 Nov. 2019.

(4) Charlet, Kate. "Understanding Federal Cybersecurity." Belfer Center. 1 Apr. 2018. Web. 1 Dec. 2019.

(5) Burgess, Christopher. "Do Cybercriminals Ever Get Extradited?" Security Boulevard. 13 Apr. 2018. Web. 1 Dec. 2019.

(6) "2016 Presidential Campaign Hacking Fast Facts." CNN, 31 Oct. 2019. https://www.cnn.com/2016/12/26/us/2016-presidential-campaign-hacking- fast-facts/index.html. Web. 25 Nov. 2019.



BIG DATA AND PRIVACY

BY: ARI WEINBERG

Every single day, petabytes of data are collected about millions of people and are stored, analyzed, and traded for profit. This new phenomenon is called big data and is expected to be a $118.52 Billion dollar industry by 2022 (1). All this data collection, however, gives rise to many questions regarding user privacy.

Before we dive into that, let's first clarify what big data is. Big data is the collection of massive amounts of data from many users for the purposes of analyzation. This data is collected from many different sources, such as online shopping activity, social media activity, web history, and government records. This data is collected both by traditional companies, who gather data from the users of their platforms, or data brokers, who collect data from multiple sources in order to aggregate it all together and sell it on for profit. This data is then analyzed for the purpose of spotting trends that can be used by companies to make better business decisions, such as what products should be developed, where advertising should be focused, etc. While big data sounds like a great resource for companies, it has many important side effects. One of the biggest side effects is the violation of privacy of the people whose data is being collected.

One of the most obvious privacy concerns is data breaches. These days, nary a week goes by that there isn't news of some big corporation being hacked and its users' personal information exposed (2). If companies are hoarding large stores of data, then they are responsible to protect that data from cyber criminals who are out to steal it and potentially leak it to the public. Norton reported that there were 3,800 publicly disclosed data breaches in the first half of 2019, with 4.1 billion record being exposed 2. These records can contain sensitive personal data, including social security numbers, passwords, addresses, and more. Data breaches are so serious that laws and regulations have been put in place that require companies to take specific steps in the event of a data breach. Many states also require companies to notify users if their personal data has been compromised.

Another big data privacy concern is deanonymization. When asked about the privacy of their users, many companies will respond that they "scrub" their data, meaning they remove all information that can be tracked back to an individual person. However, it has been shown that this is not necessarily enough. MIT scientists have shown that through a process called deanonymization, or data reidentification, data can be combined with data from other sources in order to find which individual that the data belongs to (3). In this study MIT researchers obtained location stamps from two data sets in Singapore, one from mobile phones, and the other from transit trips. They then used an algorithm to match overlapping data and were able to identify who took which trips with a 95% success rate with just 11 weeks of data. While the MIT study was not trying to unveil any personal information, it shows that bad actors can easily match anonymized data sets with personal ones, in order to unmask people's private information.

A third privacy related problem with big data is discrimination. Many people who use online services realize that they are giving up personal information and data in exchange for using the service in question. They realize that this data is going to be analyzed by big corporations, and the knowledge gained by the analyzation is going to be used to further the company's interest. What many people don't realize is that the algorithms used by big companies may unintentionally discriminate against certain groups or minorities. For example, credit companies decide whether to issue credit based on a credit score. More and more, these companies are relying on big data in order to decide that credit score, and the sources from which the data comes may be numerous. If the sources of the data in question do not equally represent all peoples and groups, it's possible that the outcomes deduced from this data may be skewed. Furthermore, the algorithm that processes this data is written by humans, who must decide which data points to include or exclude in order to give the algorithm the best chance of making an accurate prediction. It is entirely possible that the people who coded the algorithm let their biases and discriminations affect the algorithms conclusions unintentionally. The outcome of all this is the discrimination of certain people who will receive a lower credit score than they deserve, thus depriving them of opportunities such as loans or credit cards.

To conclude, in today's day and age, it is impossible to use the internet without many different companies collecting diverse data about you. Companies store this data for a very long time and use it to make predictions in order to further their interests. This process of data gathering has its benefits for both the consumer, usually in the form of free usage of many services, and companies, usually in the form of marketing predictions. However, all this data collection and sharing raises privacy concerns for those whose data is being collected and held. These privacy concerns include data breaches, data deanonymization, and data discrimination. For those who want to learn how much data is being collected about them (and I encourage you to do so), many big companies offer the ability to download all the information they have on you from their website. Apple's URL is privacy.apple.com, Google's: takeout.google.com, and Microsoft's: accounts.microsoft.com/account/privacy.

WORKS CITED

(1) MarketWatch. Big Data Analytics Market 2018 Global Size Share Growth Opportunities and Industry Forecast by Type Price Regions Key Players Trends and Demand by 2023, MarketWatch, 27 Aug. 2018, https://www.marketwatch.com/press-release/big-data-analytics-market-2018-global-size-share-growth-opportunities-and-industry-forecast-by-type-price-regions-key-players-trends-and-demand-by-2023-2018-08-27.

(2) Rafter, Dan. "2019 Data Breaches: 4 Billion Records Breached So Far." Norton, us.norton.com/internetsecurity-emerging-threats-2019-data-breaches.html.

(3) Campbell-Dollaghan, Kelsey. "Sorry, Your Data Can Still Be Identified Even If It's Anonymized." Fast Company, Fast Company, 10 Dec. 2018, www.fastcompany.com/90278465/sorry-your-data-can-still-be-identified-even-its-anonymized.