Big Data and its Implications for Privacy
(STS Topic)

Measuring Airport Similarity to Create a Towering Decision Aid
(Technical Topic)

A Thesis Prospectus in STS 4500

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

By
Michael Quinn
12/4/2019

Technical Project Team Members:
Austin Anderson
Toby Hansford
Mason Jordan
Sragi Khakurel
Chris Marshall
Katherine Taylor
Amy Xie

On my honor as a University student, I have neither given nor received unauthorized aid on this
assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signature:_*Michael Quinn*_____ Date 05/01/2020
     Michael Quinn

Approved:_____ Date _____
     Michael Gorman, Department of Engineering and Society

Approved:_____Date_____
     Cody Fleming, Department of Engineering Systems and Environment

I.   Introduction

The topic I chose to write about for my STS Research paper is on Big Data and its role in the music industry. I plan on implementing a framework of normalized deviance in order to assess how companies have grown increasingly more knowledgeable of their customers, and how this acquisition of private information is becoming normalized. There are a multitude of different music streaming platforms nowadays for listeners to choose from, with arguably the most popular being Spotify. Spotify has its own machine learning algorithms that are used to develop suggested playlists for its users ("Spotify," 2019). My STS topic will target how media giants, particularly Spotify, implement machine learning and data analysis techniques to cater to its users, and how these techniques do or do not obey ethical standards and privacy rights. To this end, their data collection and sharing methods, as well as their policy changes, will be analyzed to illustrate the ethicality of them and how it may benefit or harm the user. Specifically, I plan on discussing and analyzing the implications of these actions with regard to ethics and basic human privacy rights. I then plan on comparing the actions taken and legislation implemented by different countries and governments to protect their citizens from privacy breaches through the use of data collection. I chose this topic because I am extremely passionate for music and have a fond interest in the growing phenomenon known as Big Data and its technological and societal implications.

My technical report, on the other hand, is related strictly to my capstone project. My capstone project revolves around quantifying the benefit of installing an Air Traffic Control Tower (ATCT) at an airport. The benefits we are analyzing range from economic and financial to safety and efficiency of flights. There is a broad range of metrics to include in this analysis, and we are working with our client, the Fort Hill Group, to define said metrics. The ultimate

deliverable that the client wants us to present to them is some form of a quantitative model that can outline the benefit of installing a tower in a currently non-towered airport, and decide whether or not it is ideal. We are working to gather sufficient datasets and devise modeling techniques that will incorporate each of our metrics and weigh them appropriately in our analysis.

My STS Research topic and Technical report topic are completely unrelated. For the remainder of my prospectus, I will discuss my STS Research topic as well as the framework for which I intend on incorporating.

II.　Big Data and Big Concerns

Big Data is one of the most important and controversial topics in technology nowadays. Whether people know it or not, they are being surveyed on nearly every website, app, and device they use. With the recent surge in technology use, just over the last few decades, this phenomenon is more present than ever. The exponential growth in data generation through all types of technological portals is helping companies and firms learn an astonishing amount of information about the general public, and the related concerns regarding ethics and privacy rights are growing just as fast (Jurkiewicz, 2018).

Big Data is generally defined using three concepts; volume, velocity, and variety ("What Is Big Data," 2013). These concepts are used to describe the large amount of information (volume), the speed at which this information can be processed and analyzed (velocity), and the different types and structures of the data itself (variety). Because of the tremendous amounts of data that is generated through these three different dimensions, new data analysis techniques have been invented simply in order to process Big Data. According to a study by IBM, "users

create 2.5 quintillion bytes of data every day", meaning that 90% of the data in the world has been generated in just the last two years ("What Is Big Data," 2013). Thus, companies and industries that have access to this data hire contractors or employees that can devise coding methods for reading and manipulating the data to their advantage. These coding methods range from machine learning, predictive modeling, data mining, and more. All of these methods are used to discover patterns in data so that predictions can be made about specific user groups. Simply downloading an app, such as Facebook or Instagram, means the user is allowing these companies to gather vast amounts of information right from their phone so that they can then apply these techniques in order to cater to the user. As technological knowledge increases and more people allow companies access to their information, it is imperative that the ethical implications of these activities are prioritized.

Given the enormous amount of information that is being generated by companies in all types of industries, it is only obvious how issues regarding privacy have arisen in recent years. One of the most prominent news stories regarding data acquisition and manipulation is the Facebook-Cambridge Analytica scandal. In 2013, a researcher from Cambridge University by the name of Aleksandr Kogan created a quiz on Facebook that gathered data about the user as well as the friends of the users (Chaykowski, n.d.). Later on, it was discovered that Kogan had violated Facebook privacy policies by sharing the data with a firm known as Cambridge Analytica. Kogan was thus banned from Facebook and Cambridge Analytica was required to delete all of the data it had received. However, there are still suspicions that the company did not do everything that Facebook asked of them, despite claiming that it deleted all the information. Since then, Facebook has made many changes in both its privacy policy and user interface in order to warn users, audit third-party apps and developers, and prevent such mistakes from

happening again (Meredith, 2018). Despite the efforts made, it was alleged that Cambridge Analytica utilized the data to promote its own political agenda, namely Trump's 2016 victory, Ted Cruz's political campaign, and the Brexit vote. This is a prime example of how Big Data, gathered through a social media supergiant like Facebook, can pose many threats to the general public, in this case a political dilemma. This goes to show that in any industry, not just social media, privacy policies and regulations must be of utmost importance when using digital platforms and devices, especially as this technology is rapidly growing.

As new apps are being developed every day and companies are becoming more tech-savvy, the ethics with which organizations go about managing and utilizing their data is becoming more and more of a concern. Due to the exponential growth in both data and technology, it is crucial that the protection of privacy rights is diligently upheld.

III.   Data Analysis in the Music Industry

For my thesis topic, I decided to focus specifically on data analysis in the music industry. Apps like Spotify, TIDAL, and Apple Music are growing ever so popular, and as a result, more and more Big Data is being generated. As companies like these continue to gather more information on their users, it is apparent how the access and manipulation of private information is becoming more and more normalized. First, I will discuss the techniques that Spotify uses to create its listeners' experience, then I will describe the controversy and issues it has faced as a result of doing so.

Since its launch in 2008, Spotify has acquired 17 different companies, many of which have aided in its artificial intelligence and machine learning algorithms to help power their music discovery features (*Form F-1*, n.d.). Its most popular discover feature is its Discover Weekly

playlist, a playlist introduced every week comprised of thirty songs that it believes its user will enjoy. There are three different models that Spotify fuses to create such a playlist: collaborative filtering, natural language processing (NLP), and raw audio models (Ciocca, 2018). Collaborative filtering is used by other big companies too, like Netflix, and it works by comparing lists, or vectors, of users' most-listened-to songs and then recommending songs based on similarities. NLP is a method that involves searching the web for adjectives that are often associated with artists and songs, and using that information to recommend songs to users they think they will enjoy. Raw audio models, in short are a way of representing songs in the form of frequencies and audio frames, and then recommending songs with similar models. In conjunction, these three techniques comprise the science behind Spotify's famous Discover Weekly playlist, as well as a few other playlists such as Release Radar and Daily Mix. Today, these playlists comprise roughly a third of all user playback on Spotify, which is a huge increase in just the last few years (*Form F-1*, n.d.). The cutting-edge technology in data analysis that Spotify implements is what makes it the most popular music streaming service worldwide (Field, 2018).

Like any company, Spotify has to update and maintain its terms of service and privacy policy in order to best educate its users on what they are getting involved with. In 2015, a new policy change allowed Spotify to access users' photos, contacts, location, and more (Mason, 2015). This sparked an outcry, and although the company later added an opt-out alternative, users were furious. The problem here was that it initially did not give users an explicit option to reject these privacy breaches, meaning they had little choice but to accept the update. Soon after the release, Spotify issued an apology and clarified that users would have the ability to decide if and when Spotify could access this information. Although this did not lead to any scandals or

privacy infringements, it goes to show how quickly a supergiant like Spotify can go about gathering billions of data points about its users in the matter of a day. Aside from the dangers, though, it's important to understand that Spotify was doing this in order to enhance its users' experience and improve the app overall.

Here is where the dilemma lies: At what point do the benefits of accessing and commercializing personal information outweigh the potential harms and privacy breaches that this data acquisition poses? Of course this knowledge can help companies adjust their apps and technology to best suit the users, but sufficient action must be taken to ensure that the users are aware of and consent to the collection of the information they are supplying. As the wealth of information on the general public continues to expand, people must pay close attention to the companies and apps that they decide to share their information with.

IV.   Conclusion

Big Data is one of the most powerful tools around today. As life continue to become more digitized, and technology continues to swarm the globe, the potentials that this application holds seem to be growing astronomically. As with any new invention, Big Data hosts a wide array of benefits and dangers that it can bring about. What is most important in this day and age, is that users become conscious of this and learn how to protect themselves from malicious attacks. As part of my thesis, I plan to explore avenues that may be taken in order to protect and educate people on the privacy matters of their shared information. This includes the possible creation of third-party apps to summarize privacy policies for users, as well as legislation passed to protect the private information of citizens, such as the EU's General Data Protection

Regulation (GDPR). Until the power of Big Data is harnessed, supervised, and utilized for the overall benefit of society, the threats it poses cannot be taken lightly.

References

Chaykowski, K. (n.d.). *Mark Zuckerberg Addresses "Breach Of Trust" In Facebook User Data Crisis*. Forbes. Retrieved April 4, 2020, from https://www.forbes.com/sites/kathleenchaykowski/2018/03/21/mark-zuckerberg-addresses-breach-of-trust-in-facebook-user-data-crisis/

Ciocca, S. (2018, April 5). *How Does Spotify Know You So Well?* Medium. https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe

Field, M. (2018, February 22). The best music streaming services: Apple Music, Spotify, YouTube Music and Amazon Music compared. *The Telegraph*. https://www.telegraph.co.uk/technology/0/best-music-streaming-services-apple-music-spotify-amazon-music/

Form F-1. (n.d.). Retrieved April 30, 2020, from https://www.sec.gov/Archives/edgar/data/1639920/000119312518063434/d494294df1.htm#rom494294_4

Jurkiewicz, C. L. (2018). Big Data, Big Concerns: Ethics in the Digital Age. *Public Integrity*, *20*(sup1), S46–S59. https://doi.org/10.1080/10999922.2018.1448218

Mason, P. (2015, August 23). The Spotify privacy backlash: What is my personal data really worth? | Paul Mason. *The Guardian*. https://www.theguardian.com/commentisfree/2015/aug/23/the-spotify-privacy-backlash-what-is-my-personal-data-really-worth

Meredith, S. (2018, March 21). *Here's everything you need to know about the Cambridge Analytica scandal*. CNBC. https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html

Spotify: Big Data Shows Big Results. (2019, October 4). *Northeastern University Graduate Programs*. https://www.northeastern.edu/graduate/blog/spotify-big-data/

What is Big Data. (2013, June 4). *Villanova University*. https://www.villanovau.com/resources/bi/what-is-big-data/