Big Data and its Implications for Privacy

A Research Paper submitted to the Department of Engineering and Society

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

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Michael Quinn
Spring, 2020

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

I. Introduction

This research paper is primarily concerned with the relatively new and growing phenomenon known as Big Data, what role it plays in the modern world, and the problems, mostly of the ethical nature, associated with it. Part of the research will explore, in particular, the music industry and how Big Data has shaped it. The nature of Big Data is immense and profound, and its manipulation can have very positive or negative effects on the general public. With such a vast wealth of digital information available at the fingertips of public and private entities, individuals must be cautious of the applications they use and the information they share in order to avoid privacy exploitations. In addition to educating the public, legislation and regulations should be put in place to protect citizens and their personal information. I plan on incorporating an ethical framework to analyze the issues concerning privacy that have surfaced as a result of Big Data.

II. Big Data

Big Data is arguably one of the most impactful and relevant developments in data analysis, and the digital world, to this day. Its origin can be traced back to before the year 2000, where most data scientists attribute the term's founding to John Mashey and his colleagues at Silicon Graphics when they produced a slide deck entitled "Big Data and the Next Wave of InfraStress" for a technical seminar in 1998 (Diebold, 2012). Since then, the term has gained increasing popularity along with the recent explosion of smartphones, social media, and the Internet. But it is just as important that, along with the increasing use of these sprouting technologies, the ethics behind these technologies is closely monitored and regulated.

The National Institute of Technology and Standards (NIST) defines Big Data as "extensive datasets… that require a scalable architecture for efficient storage, manipulation, and analysis, " (NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, 2015). They describe it as a drastic, one-time shift in the fundamental architecture for efficiently handling the massive amounts of data that are present today. A major component of this shift is transitioning from a vertical database system scale to a horizontal one. Vertical scaling refers to upgrading or downgrading the actual hardware of a server to either increase capacity or decrease capacity, respectively (Kuhlenkamp et al., 2014). The latter is primarily used to save money. The focus of Big Data is horizontal scalability, which involves integrating many individual nodes or resources together to act as a cohesive system. This presents an advantage over vertical scaling because it avoids hardware limitations and instead uses a system set in parallel to increase or decrease computing power. Aside from its definition, Big Data can be summarized by using four different words. My prospectus included the three V's: volume, velocity, and variety, which were initially used by Doug Laney in 2001 (Diebold, 2012). The NIST, however, includes a fourth V word: Variability. This fourth V is similar to velocity, but it refers to the change of data over time, whether it be the flow rate, size, format, or something of the like. This can be thought of as a graph, with peaks and troughs that resemble periodic spikes and drops in the rate of data being collected and analyzed at different times, perhaps due to a current event or trend on social media. There are many ways to describe the term Big Data, but in order to gain a more holistic understanding, one must survey its associated technologies.

There have been many new technologies created to accommodate for this immense and growing phenomenon known as Big Data. The sheer number of data bytes that are being produced nowadays has called for a revolution in the technologies used to aggregate and analyze

data. One of the most popular ones is cloud computing. Cloud computing is a concept and service that allows storing and accessing data as well as computing power on remote computing resources that are accessed via the Internet, rather than on a local hard drive or computer

Figure 1. The interior of an Amazon data center where tens of thousands of servers work around the clock to store information (Cole, G., 2014, September 24).

(Manyika et al., 2011). This technology is used by many businesses to store their clients' or their own data and protected information in safe reserves. A number of benefits come along with cloud computing, such as more cost-efficient IT systems, easier scalability, information security and backups, efficient collaboration and sharing, along with many other advantages (Avram, 2014). Another popular technology that has arisen along with the surge of Big Data is data warehouses. Data warehouses are a sort of environment that allow for the facilitation of large-scale management and storage of Big Data (Santoso & Yulia, 2017). These are predominantly used by big businesses and enterprises and are key in decision-making processes and analytical

reports. They are a fundamental component of what is known as Business Intelligence, which will be explained next. Business Intelligence is comprised of all the tools, technologies, and softwares that businesses use to translate data into actionable insights that is used to monitor the status of the business (Fruhlinger, 2019). It involves manipulating the data that is stored in data warehouses to create visualizations, graphs, dashboards, and more to display detailed findings and paint a picture for the end-user. One of the most popular Business Intelligence softwares is Microsoft Power BI, a service created by Microsoft that is used by companies such as Adobe, Kraft Heinz, Aston Martin, and more. Through the synthesis of cloud computing, data warehouses, and other unnamed technologies, a successful Business Intelligence model can utilize the massive amounts of Big Data to optimize a business's operations and achieve their goals.

Another result of the Big Data boom is a revolution of coding techniques to extract and study the large amounts of data being collected. One of most popular techniques is known as machine learning. Machine learning, which I briefly mentioned in my prospectus, is a form of artificial intelligence that involves the development and refinement of algorithms used by computers to make complex and intelligent decisions (Manyika et al., 2011). Computers are programmed to constantly reevaluate and improve their coding algorithms through empirical analysis of the data at hand in order to make decisions that realistically only a human mind could carry out. Natural Language Processing (NLP), which was discussed in my prospectus, is an example of a machine learning algorithm. Data Mining, which I also mentioned in my prospectus, is another computer science process that has changed the way Big Data can be manipulated. Data mining is a means of discovering important or relevant patterns and relationships within large sets of data (Clifton, 2019). Companies will use data mining to find

trends and general knowledge about their client base in order to improve their business practices and marketing strategies. There are many techniques that are specific to data mining, one being predictive modeling which was aforementioned in my prospectus. Predictive Modeling is the process of using sample data to train a computer in estimating a particular value or outcome of an event. This form of data mining is useful for when businesses want to survey a prototype or predict the success of a product. However, there are hundreds, if not thousands, more coding techniques that have been put in practice by large companies and corporations with the goal of harnessing and extrapolating Big Data.

Spotify is arguably the most popular music streaming service available worldwide. This is due largely in part to its successful implementation of some of the algorithms and techniques previously described. Since its launch in 2008, Spotify has acquired 17 different companies, many of which have aided in its artificial intelligence and machine learning algorithms that power its music discovery features (*Form F-1 Registration Statement*, 2018). Its most popular feature is its Discover Weekly playlist, which is a short list of recommended songs given to each user on a weekly basis that it calculates based off the user's listening history. One might wonder, how can sounds, of all things, be analyzed by a computer and transformed into recommendations for similar music. Spotify uses three different methods to do this: collaborative filtering, natural language processing (NLP), and raw audio models (Ciocca, 2018). Collaborative filtering is a process used by other big companies too, like Netflix, that works by comparing lists, or vectors, of users' most preferred items (in this case, songs) to other users, finding matches, and then recommending songs based on similarities (Sarwar et al., 2001). 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

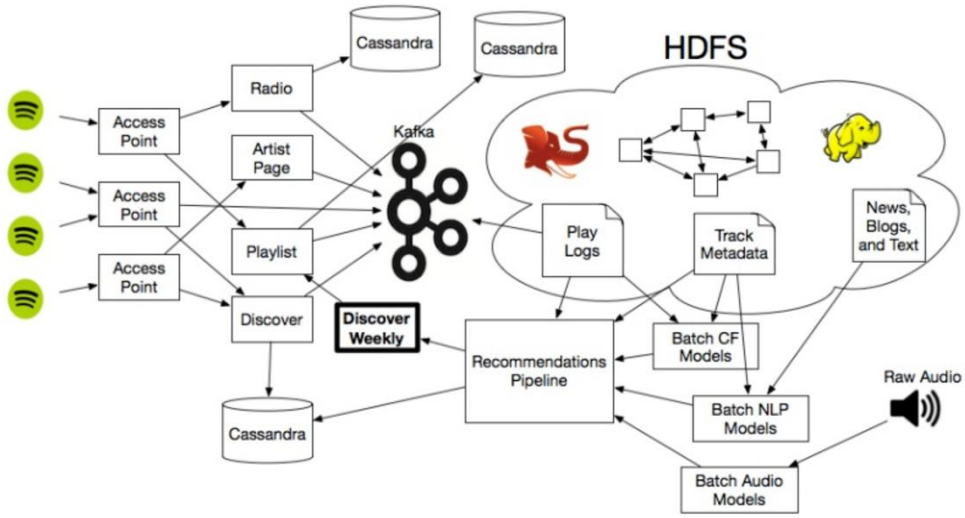

Figure 2 A flow diagram of the various processes and algorithms Spotify uses to curate the Discovery Weekly playlist for all of its users (galvanize, 2016, August 22).

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.

It is transparent that in just the last couple decades, Big Data has had drastic effects on the technologies we interact with every day. From the invention of new technologies to the construction of massive computers and warehouses, the Big Data revolution is one of the most profound technological shifts in recent time. However, the ethical implications of these new technologies must be heavily considered before they are introduced and implemented. The amount of data and information that billions of people are sharing, often unknowingly, through various apps and technologies presents a dilemma of privacy. The manipulation of Big Data can be just as easily used for malicious intent as it can be for societal benefit. It is imperative that

governments allocate sufficient resources to monitor the research behind these technologies and impose regulations to ensure the safety of the general public and avert any unethical misuses.

III. Big Data, Big Issues

Big Data poses many big concerns not only for the tech industry but for the general public as a whole. The amount of data being generated nowadays is an astonishing number, and it is predicted to continue growing exponentially. In 2018, the International Data Corporation predicted that the amount of data in the world will grow from 33 zettabytes in 2018 to 175 zettabytes by 2025 (Reinsel et al., 2018). For reference, the average computer has about 250

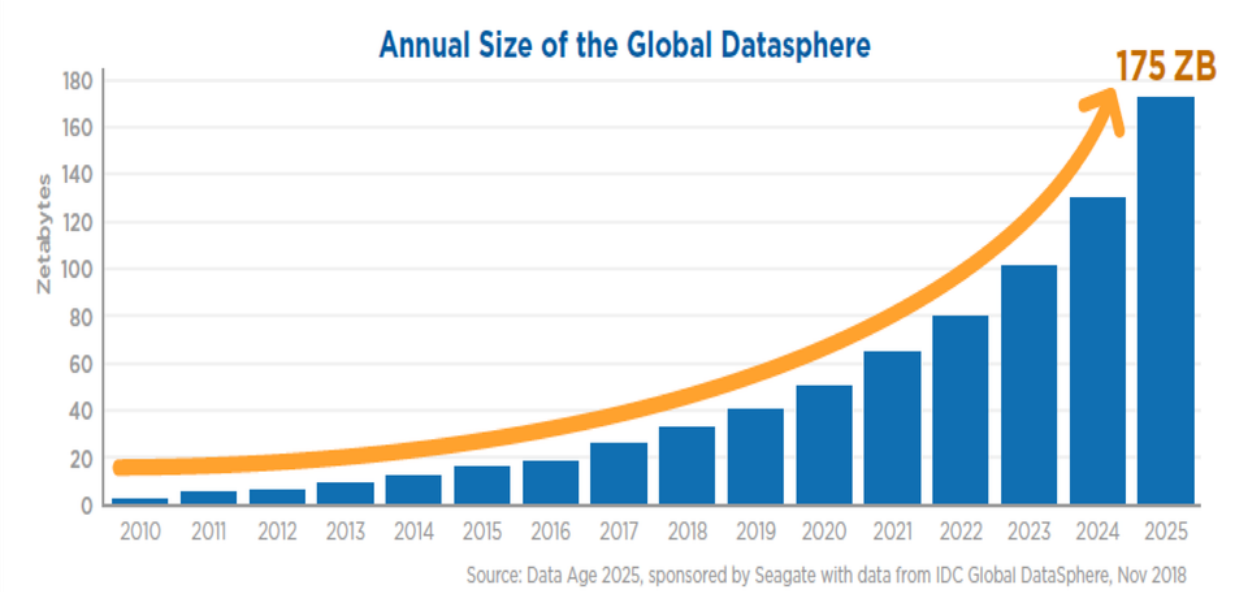

Figure 3. A graph of the historical size as well as future predicted size of the total amount of data in the world (Coughlin, 2018, November 27).

gigabytes of storage, and one zettabyte is equivalent to one trillion gigabytes. Given this large amount of data, and all the forms it takes as a result of the use of smartphones, social media, internet browsing, and more, public entities like the federal government as well as companies in the private sector have an astounding amount of information about the general public at their fingertips. On occasion, this had led to privacy breaching issues and public outcries. But most of

the time, people are not aware how much and how often they are sharing information about themselves. Guidelines and regulations must be put in effect at the federal level to protect the general public and prevent the exploitation of innocent individuals.

Essentially everyone that has used the Internet has shared their data online. Search engines, entertainment services, and social media giants like Google, Facebook, Netflix, and YouTube have a surprising level of knowledge on its users; from physical location, social relationships, hobbies and interests, personal photos, and more, these companies have a massive array of data points that they process and analyze to cater to their customers and enhance the user experience. However, some information shared with these corporations are private and must be securely stored, like street addresses, credit card information, and social security numbers. So how do these companies make sure they aren't breaching privacy rights, and how do they protect the private information from hackers? One of the most expensive, news-breaking privacy incidents occurred in 2016, when Uber revealed that hackers had stolen personal information (names, email addresses, phone numbers, and even driver license numbers) from over 50 million of its users and drivers (Chappell, 2018). The first issue here is that Uber failed to protect this wealth of personal information. This is a result of the lack of knowledge, technological expertise, and educated individuals needed to successfully handle and guard these large amounts of data from malicious attackers. However, the reason Uber paid nearly 150 million dollars in settlement claims is because instead of reporting the security breach, which is mandated by law, they paid off the hackers for $100,000. This is the second issue at hand. The public puts a certain level of trust in these corporations when they share their information. When companies choose to act illicitly, it jeopardizes, in this case, millions of people and directly violates their privacy rights. The U.S. government must act ethically and do so by funding research on this subject in order to

devise and implement appropriate legislation that ensures sufficient oversight on the wealth of data shared by its citizens on the Internet.

Just like every other tech-related industry, the music industry has its own concerns due to its use and sharing of user data. Back in 2015, Spotify updated its privacy policy to reveal that it is able to collect certain user information such as location, photos, contacts, and more (Faughnder, 2015). This sparked an outcry among the customer base and many users actually decided to discontinue their subscriptions. Days later, CEO Daniel Ek came out to clarify, what some described as the "creepy" policy changes. He noted that users have the option to share certain information with Spotify, and that the company was going to revise the policy language to be more transparent. As of the beginning of 2020, Spotify's privacy policy states that, from simply using the service, the company collects data such as cookies, IP addresses, among other online identifiers, network connection type, non-precise location, motion-generated data through the use of accelerometers, and much more (*Spotify Privacy Policy*, 2020). Furthermore, upon approval of the user, Spotify has the ability to collect more personal information, such as voice and photo data. They then partner and share these various data types with third parties to gauge certain interests, target ads, engage in research studies, and more. Spotify is just one of many high-profile companies that has access to these kinds of information and shares it with other companies to enhance user experience while also optimizing business operations. As of now, according to federal law, this is a completely ethical practice. While this data collection may not disturb some people, a considerable portion of the world's Internet surfers and app users are just entirely unaware of it.

Not only is there an absence of rigid laws and guidelines to protect the average American's private information, many Americans are simply uneducated on the issue. In order to

protect American citizens, educate the public, and keep pace with leading nations, the federal government needs to work to administer national safety guidelines that corporations are mandated to observe and obey.

IV. Governments in Action

Nations all around the world have already imposed national laws that govern the acquisition and use of personal data by businesses. Japan has the Act on Protection of Personal Information (APPI), Brazil has *Lei Geral de Proteçao de Dados* (LGPD, "General Data Protection Act"), and the EU and its 27 members have the General Data Protection Regulation (GDPR), which the LGPD is closely modeled after. These are just some of the world's top

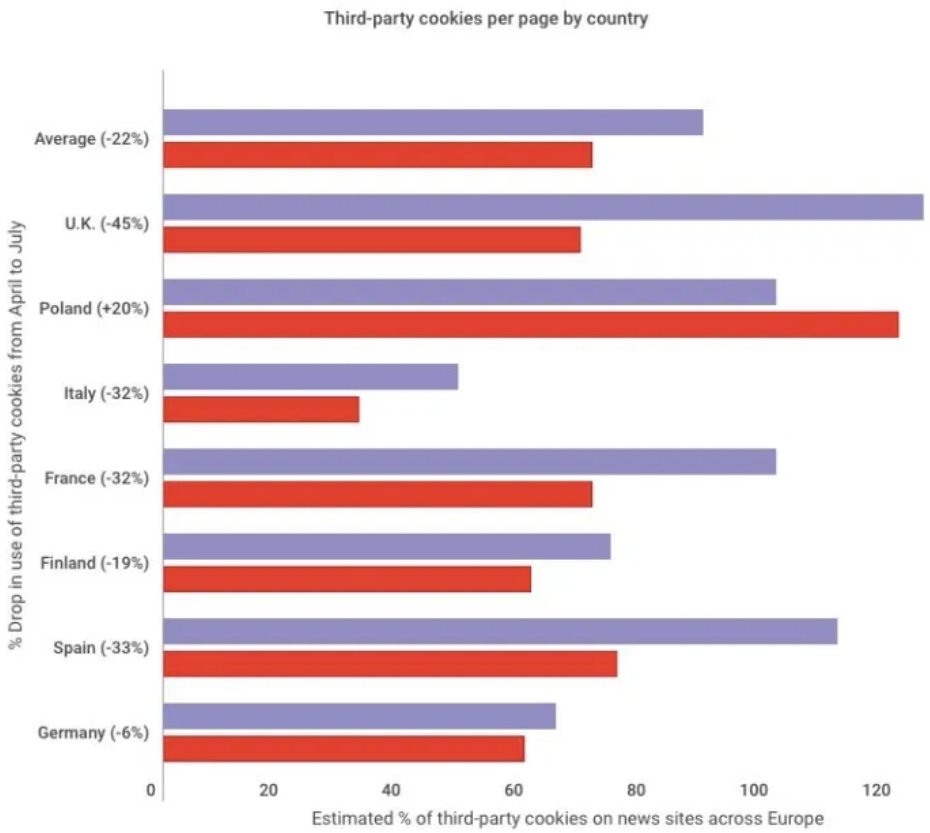

Figure 4. A chart of the percent difference in cookies used by third parties per web page for multiple countries in the EU before and after the GDPR was implemented, with an average 22% drop among all countries. Cookies are small pieces of data stored on websites when a user browses (Davies, 2018 August 24).

countries that have taken the initiative in imposing national data privacy regulations to protect their citizens. However, despite the fact the U.S. is constantly considered one of the most advanced nations in the world, it is yet to implement any legislation that is comparable to the acts mentioned above.

The U.S. lacks a singular, comprehensive law for protection and regulation of the personal data of its citizens. Right now, the only relevant laws in place are various statutes enforced by and for each state, along with some sector-specific laws that concern particular types of data (Chabinsky & Pittman, 2019). On the other hand, the GDPR is "the toughest privacy and security law in the world", that protects all citizens within the EU and applies to any business or organization that collects or processes their data, regardless of location (*GDPR Archives,* n.d.). Entities that process data are required to abide by seven protection principles, which include lawfulness, fairness, and transparency to the data subject, accuracy of personal data, storage limitations, and other rules. The GDPR also includes specific guidelines for what defines obtaining consent from a data subject, which is crucial in the notification and education of citizens in regard to the sharing of their personal information. This sort of formulation is what the U.S. government must work to impose in order to protect its citizens for the future to come, as Big Data and its manipulation by corporate entities will only continue to expand.

As with any law, the protection and welfare of the general public must be of top priority. The ethical implications that coincide with the use of Big Data are vast, and must be considered as the basis for any legislation that is to be passed in conjunction with this technology. As the global wealth of data continues to increase at an exponential rate, and companies work to invent and refine new ways of accessing and processing personal data, it is urgent that the U.S. government implements a data protection and regulation act at the federal level.

V. Conclusion

Big Data is at the forefront of analytics, both for commercial and educational use. As the knowledge of it as well as the data itself continues to expand at an increasing rate, governments that haven't already must act to educate and protect its constituents from being exploited by data collectors and processors. The general public puts its trust in businesses to secure and not misuse the information they provide them with, and this is a trust that must remain intact for the benefit and welfare of society. Aside from private corporations that have access to personal data, there are straight up malicious attackers and hackers that pose as direct threats to innocent civilians. Not only must data-collecting entities act ethically and keep their promise to the general public that they will not commit misconduct, they must learn how to protect these massive collections of information from said hackers. The issue at hand is highly volatile, and in an age of expanding knowledge and technologies, these attackers are more elusive than ever. Governments must act soon and accordingly in order to safely harness the immense power of Big Data, protect its citizens, and prevent an unethical catastrophe.

References

Avram, M. G. (2014). Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. *Procedia Technology*, *12*, 529–534. https://doi.org/10.1016/j.protcy.2013.12.525

Chabinsky, S., & Pittman, P. (2019, March 7). *International Comparative Legal Guides—USA: Data Protection 2019* [Text]. International Comparative Legal Guides International Business Reports. https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa

Chappell, B. (2018, September 27). *Uber Pays $148 Million Over Yearlong Cover-Up Of Data Breach*. NPR.Org. https://www.npr.org/2018/09/27/652119109/uber-pays-148-million-over-year-long-cover-up-of-data-breach

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

Clifton, C. (2019, December 20). *Data mining | computer science*. Encyclopedia Britannica. https://www.britannica.com/technology/data-mining

Cole, G. (2014, September 24). *Amazon May Build Data Center in Ohio—Avalara*. Taxrates. https://www.avalara.com/taxrates/en/blog/2014/09/amazon-may-build-data-center-ohio.html

Coughlin, T. (2018, November 27). *175 Zettabytes By 2025*. Forbes. https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/

Davies, J. (2018, August 24). The impact of GDPR, in 5 charts. *Digiday*. https://digiday.com/media/impact-gdpr-5-charts/

Diebold, F. X. (2012). On the Origin(s) and Development of the Term “Big Data.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2152421

Faughnder, R. (2015, August 22). *Spotify CEO Daniel Ek on privacy policy dust-up: “Sorry.”* Los Angeles Times. https://www.latimes.com/entertainment/envelope/cotown/la-et-ct-spotify-ceo-privacy-policy-sorry-20150821-story.html

Form F-1 Registration Statement. (2018). https://www.sec.gov/Archives/edgar/data/1639920/000119312518063434/d494294df1.htm#rom494294_4

Fruhlinger, M. K. P. and J. (2019, October 16). *What is business intelligence? Turning data into business insights*. CIO. https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html

galvanize. (2016, August 22). Ever Wonder How Spotify Discover Weekly Works? Data Science. *Galvanize Blog*. https://blog.galvanize.com/spotify-discover-weekly-data-science/

GDPR Archives. (n.d.). GDPR.Eu. Retrieved April 6, 2020, from https://gdpr.eu/tag/gdpr/

Kuhlenkamp, J., Klems, M., & Röss, O. (2014). Benchmarking scalability and elasticity of distributed database systems. *Proceedings of the VLDB Endowment*, *7*(12), 1219–1230. https://doi.org/10.14778/2732977.2732995

Manyika, J., Chui, M., Institute, M. G., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey.

NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. (2015). *NIST Big Data Interoperability Framework: Volume 1, Definitions* (NIST SP 1500-1; p. NIST SP

1500-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1500-1

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World from Edge to Core*. 28.

Santoso, L. W., & Yulia. (2017). Data Warehouse with Big Data Technology for Higher Education. *Procedia Computer Science*, *124*, 93–99. https://doi.org/10.1016/j.procs.2017.12.134

Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the Tenth International Conference on World Wide Web - WWW '01*, 285–295. https://doi.org/10.1145/371920.372071

Spotify Privacy Policy. (2020, January 1). Spotify. https://www.spotify.com/us/legal/privacy-policy/