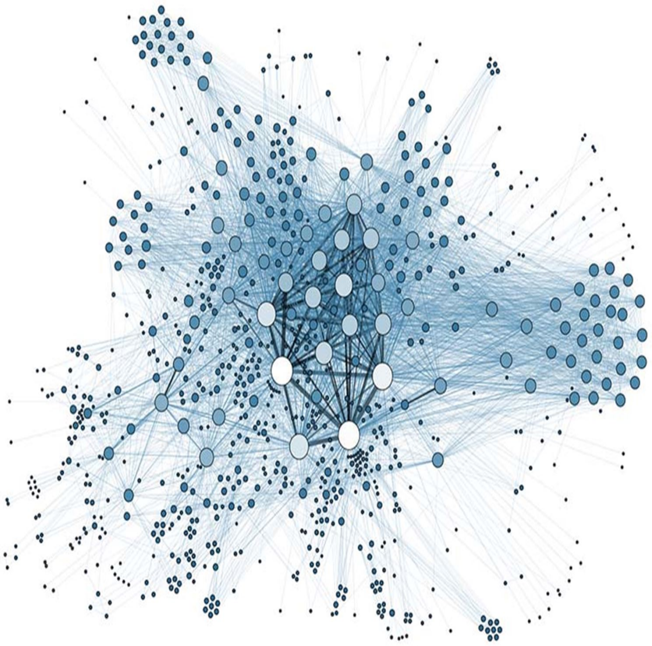
Predictive Analytics: From Past to Present

An accessible white paper for university students
Submitted to Dr. Mark Spielmacher
April 9, 2019



BY TEAM 42:
FINN PLUMMER
JOSHUA GOLDSMITH
BRENNEN CREIGHTON-YOUNG

Executive Summary

The report was written to raise awareness of the influences and uses of predictive analytics in the university age group. Many students are unaware of how predictive analytics is being used around them everyday and an uncertainty of predictive analytics leads to disadvantages in the students future careers. It is also important for students to understand how their data is being used in predictive analytics.

The origins of predictive analytics has deep roots in the insurance industry, since the 1680's predictive analytics showed how powerful of a tool it is. When computers were introduced the efficiency of predictive analytics increased substantially and the uses became more broad. Predictive analytics showed its strengths alongside computers with the Manhattan Project, MIT's differential analyzer and IBM's punched card machines. With the introduction of big data, heterogeneous data analysis and inference based methodology, predictive analytics became increasingly popular and now a staggering percentage of businesses owners say that predictive analytics provide values everyday within their organization. With such an increase in popularity it has ethical problems regarding privacy and how data is collected. Another issue is that predictive analytics creates problematic feedback loops whether it is intended to or not.

It is clear that predictive analytics is being used everywhere, and that predictive analytics has became a mainstream tool for businesses. Students in the university age group have heard the concept of predictive analytics before, however the students do not to what extent predictive analytics is being used around them. It is important for students at this age going into careers were predictive analytics will be even more prevalent know what it is so they can understand when and how it is being used.

Predictive Analytics: From Past to Present

By Team 42:
Finn Plummer, Joshua Goldsmith and Brennen Creighton-Young

Contents

1 Introduction 1

2 Methodology 2

3 History 2
3.1 The Very Beginning 2
3.2 Evolving Alongside Technology . . 3
3.3 Importance of Data Mining 3
3.4 Effects of Machine Learning 3
3.5 Proceeding into the Present 3

4 Predictive Analytics Today 4
4.1 Computation in Predictive Analytics 4
4.2 Prediction as a Commercial Tool . . 5
4.3 Which Data are Used? 5

5 Societal Impacts of Predictive Analytics 6
5.1 How We Get Information 6
5.2 Public Spending 6

6 Conclusion 7

7 Works Cited 7
7.1 Written Works 7
7.2 Figures 8

8 Appendix 8

1 Introduction

Imagine if you were able to closely approximate what the value of a specific stock would be in a year from now or what the chances are that your boat will sink in the next 5 years. You would be able to prevent potential losses and reduce risk in the future. This insight is what predictive analytics strives to provide to anyone who uses it. In a non-technical sense, predictive analytics is using a large amount of data to build predictive models and answer the question, “What will come next?” This is a question that has an extremely valuable answer for your future employer, local businesses, big corporations and entire industries. Predictive analytics is providing companies and entire industries this valuable insight. It is important to understand how these companies and industries are using predictive analytics to their advantage because this information is influencing everyone’s everyday lives.

This report will be directed towards those who are in university with co-op, job and career opportunities in their near future. Focusing on the essential information on predictive analytics will provide those who are moving into this stage information in the ways predictive analytics will be influencing their life in numerous ways.

This report will start with the beginning of predictive analytics and its first uses within the insurance industry and show how predictive analytics dramatically evolved with the introduction of computers. Introducing technical concepts the report will continue the discussion on the uses of predictive analytics from the 1940's into the present day highlighting substantial uses during this time period. With such a wide use of predictive analytics ethical problems and societal impacts have occurred, the last section of the report will provide an in depth analysis of these issues.

While the question of "What will come next?" provides a non-technical definition of what predictive analytics. It is important to expand this definition to include a technical definition of predictive analytics before talking about it throughout the report. This report will use Mykoa Pechenizkiy's technical definition of predictive analytics;

Predictive analytics as a research field studies how to extract useful knowledge from various data sources to induce different kinds of predictive models, which are ubiquitous for data-driven optimization, decision support, and decision making.

(Pechenizkiy, 2015)

The report will explain the technicalities in the second section of predictive analytics further however it is important to have a definition for context and the report will refer to this definition of predictive analytics.

2 Methodology

For this project, our team used a mix of professional articles, online sources and books to obtain and deliver the best possible information available to us. We found a lot of the information through Google Scholar, the Waterloo library website and regular Google searches to find sources. We feel that the articles and websites used to gather information are of good integral quality and meet the standards for the researching of a professional paper such as ours. We primarily used these sources, to produce evidence and statistics to support our claims and ideas.

We found that the websites of Stastia and InternationWorldofStats provided substantiated statistical evidence. The statistics provided appear to be credible and trustworthy, used by multiple reliable and respected websites. Further, we used books, including "Weapons of Math Destruction", to gain a better understanding of the negative societal implications of Predictive Analytics. Other sources included research papers and scholarly article for the technical research portion of our presentation

3 History

3.1 The Very Beginning

Figure 1: *The exterior of the Lloyd's building, the headquarters of Lloyd's of London*

Underwriting as defined by the Oxford dictionary is, "to set one's name to (an insurance policy) for the purpose of thereby becoming answerable for a designated loss or damage on consideration of receiving a premium percent" and is used within modern-day insurance (Underwrite, n.d.). When someone wants to insure an something they must undergo an underwriting process that an insurance company will use to try and "accurately predict future losses and price the products that protect against those losses" (Nyce, 2007, p. 5, emphasis added). It is apparent that the use of underwriting and predictive analytics compliment each other as they are both interested in trying to figure out what will happen next. Predictive analytics has been and still is used within the underwriting process because of the direct similarities they share. The first recorded professional uses of underwriting and predictive analytics was in the 1680's through a coffee house with the name Edward Lloyd's Coffee House. The coffee house was a bustling shop that offered a place for businessmen to sell their marine insurance to the ship owners who would come in from the harbour (www.lloyds.com). The coffee shop grew into one of the main places for ship owners to buy insurance for their boats and evolved into Lloyd's, today Lloyd is a large insurance and reinsurance market. Underwriters were crucial to this success and as the popularity of insurance grew and the need for underwriters increased, with an increase of underwriters predictive analytics evolved. This marked the beginning and first uses of predictive analytics in business and from the early 1700's until the 1900's predictive analytics had limited capabilities and was used mostly within insurance.

3.2 Evolving Alongside Technology

As technology advanced and computers were introduced in the 20th century the possibilities for predictive analysis expanded. The introduction of computers allowed for data to be stored in large quantities and for mass calculations to be performed quicker. Data mining and machine learning soon became terms many people were familiar with. Data mining and machine learning both had a major impact on predictive analytics, changing the efficiency and applications of predictive analytics.

3.3 Importance of Data Mining

Figure 2: *A data mining graphic*

As previously mentioned predictive analytics was used initially without the use of computers and assistance in doing computations. With the possibilities to have a larger pool of information predictive analytics became more closely tied with data mining. Abbot states that "the algorithms and approaches are generally the same" between data mining and predictive analytics as they both are searching the past and the data, however the main contrast is what each will try to accomplish with that data (Abbot, 2014, p. 13). Data mining has been used "in a wide variety of fields, including finance, engineering, manufacturing, biotechnology, customer relationship management, and marketing," and historically predictive analytics had been contained within the insurance industry (Abbot, p. 13). When predictive analytics became able to adopt the qualities of data mining it allowed predictive analytics to also become applicable in a larger variety of fields.

3.4 Effects of Machine Learning

Figure 3: *An artificial intelligence graphic*

Predictive analytics was able to become a more powerful tool with the introduction of machine learning. As defined by Kelleher, Mac Namee & D'Arcy machine learning is "an automated process that extracts patterns from data" and machine learning became prevalent once technology had advanced to the point where it could perform these automated actions. (Kelleher et al., 2015, p. 3) Predictive analytics as defined in the introduction is building predictive models from the information that is available at hand. When machine learning was introduced to predictive analytics it allowed an efficient way to find these predictive models within data sets. Previously these predictive models were prone to human-error and took extremely long in comparison to the time it took once machine learning become used. Machine learning made predictive analytics efficient and a powerful tool to quickly build predictive models and is a major component to what predictive analytics has become.

3.5 Proceeding into the Present

Predictive analytics has the possibility of being able to tell what is going to happen in the future by knowing what has happened in the past. Given that the insurance industry relied on predictive analytics warrants how powerful of a tool it is. The improvement in efficiency and accuracy that predictive analytics was given with machine learning as well as the foundation data mining provided for predictive analytics to be able to lean on allowed predictive analytics to not be contained within insurance. Now that predictive analytics is no longer contained within insurance and has spread to become a part of everyday life, it is important to understand how it was first used to be able to see how the future of technological advancement will change predictive analytics. The intricacies of how predictive is already in our lives will be discussed further in the rest of the report and will reinforce and build on how predictive analytics has been woven into our lives.

4 Predictive Analytics Today

Predictive analytics has grown in scope and utility in in recent years, primarily due to technological advancements. Though the trend of increasing technological capabilities is clear to the public, largely due to the availability of higher-quality consumer goods and software, technological advancements pervade through industries in ways that are less immediate to the general populous. Predictive analytics presents itself as an example of this phenomenon; its humble beginning as an underwriting technique has evolved into a methodology used in industries including healthcare, advertising, humanities research and economics, which has computation at its core. Though the widespread use of predictive analytics has become ubiquitous in industry, it also brings with it deep social implications.

4.1 Computation in Predictive Analytics

The advent of mechanical computers had an impact on predictive analytics in real-time. Perhaps the most historically-rich example of the in-tandem evolution of computational ability and the use of computation for forecasting was the use of technology in the Manhattan Project, wherein tens of thousands of specialized personnel worked on developing nuclear weapons for use by the United States in World War 2. For the project, world-renowned physicists Richard Feynman and Nicholas Metropolis implemented computer simulations which were hitherto unattainable due to the lack of sufficient computational capabilities. The arrival of primitive computing machines such as MIT's differential analyzer and IBM punched-card machines "enabled scientists to accurately predict other physical scientific phenomena, such as the weapon's explosive yield, pertaining to the Trinity test of July 16, 1945" ("Punched Cards to Petaflops", 2013, p. 37). This example took place at the beginning of the era of computation in predictive analytics. And, perhaps unsurprisingly, expanding technological capacity has affected predictive analytics ever since.

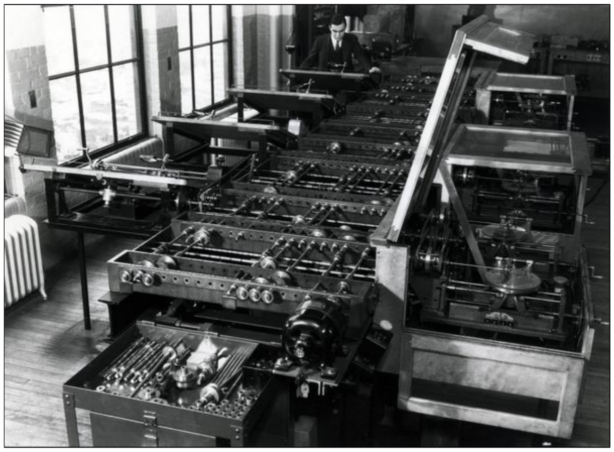


Figure 4: *A photograph of a differential analyzer*

Though statistics, which provides the technical framework for predictive analytics, is a centuries-old field of study, modern computer processing capabilities have allowed for more intricate implementations of statistical methods than ever before possible. In turn, predictive analysts have increasingly robust computational tools available to them. According to Wayne W. Eckerson (2007), evolving computer architecture has made it possible to not have to solely rely on descriptive statistical tools - the use of mean, median, mode and graphics in studying data - rather, one can now employ a breadth of techniques when performing predictive analytics. Furthermore, some of these computationally-heavy tools allow for the analysis of large, heterogeneous datasets of the sort that were historically intractable to study (Benjelloun, Lahcen & Belfkih, 2017). Due to the fundamental use of statistics in predictive analytics, these recent advancements in statistical computing capabilities suggest that predictive analytics itself has benefited from the trend of increasing processing power. In particular, predictive analytics is becoming more capable of treating complex data with increasingly versatile tools.

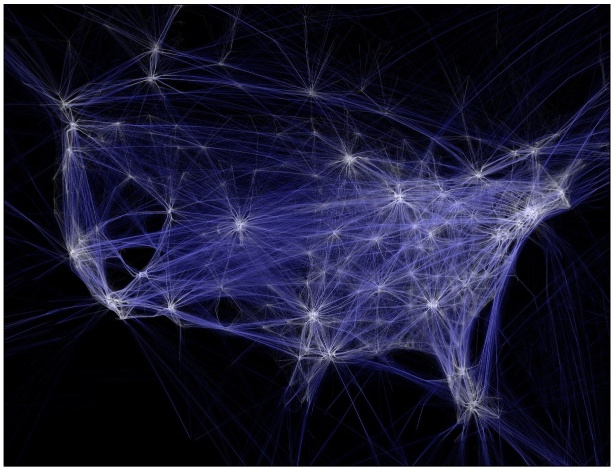


Figure 5: *A visualiazion of flight patterns - an example of complex data*

4.2 Prediction as a Commercial Tool

Along with the prevalence of technology in the modern era has come the ability for more organizations to implement predictive analytics. Furthermore, the recent progress in the analytics of big data, which refers to extraordinarily large collections of data, is contributing to the ubiquity of predictive analytics. According to Philip Russom, the director of research at Transforming Data With Intelligence, we have that "Roughly three quarters (74%) of organizations surveyed have adopted some form of analytics today, regardless of the analytic method or tool type, whether with big data or not" (Russom, 2011, p. 10) Russom goes on to describe how predictive-analytical tools, particularly those that work with very large datasets, are reasonably affordable to mid-sized organizations . Given that the execution of predictive analytics relies so heavily on computer processing, its democratization must be in response to the commercialization of more powerful computation. In addition, Eckerson (2007) presents polling data which suggest that up to 93% of organizations have had moderate-to-very high increases in business value when implementing predictive analytics. The overwhelming effectiveness of predictive analytics suggests that its prevalence has grown alongside that of technological capability and will likely continue to do so.

4.3 Which Data are Used?

While predictive analytics is about the use of data for forecasting, big data technologies facilitated its ability to make connections between exceptionally obscure sets of data. This is done with data mining, a technique used to distill meaningful information from enormous sets of heterogeneous data. When applied to online user data this technique allows for inferences to be made about the traits of individual users. Understandably, this has presented itself as the primary ethical concern currently surrounding predictive analytics. Stanford researchers mention that "online services routinely mine user data to predict user preferences, make recommendations, and place targeted ads. Recent research has demonstrated that several private user attributes (such as political affiliation, sexual orientation, and gender) can be inferred from such data" (Ionanidis et al., 2014, p. 1), and thus are actively researching the security of various prediction methods with regards to privacy. To further the issue, the big data mining status quo circumvents regulations that were introduced to limit privacy exploitation:

> Unlike previous computational models that exploited known sources of personally identifiable information ("PII") directly, such as behavioral targeting, Big Data has radically expanded the range of data that can be personally identifying. By primarily analyzing metadata, such as a set of predictive and aggregated findings, or by combining previously discrete data sets, Big Data approaches are not only able to manufacture novel PII, but often do so outside the purview of current privacy protections. (Crawford & Schultz, 2014, p. 93)

Evidently, the potential for corporations to infer qualities of their users without their consent constitutes an ethical problem. Furthermore, the capability of predictive analysis techniques to breach legal privacy agreements is problematic. Lest we forfeit our personal confidentiality protections, it is important that we, as a society, take a closer look at predictive analytics as a whole. See the next section for a further discussion of the societal impact of predictive analytics.

5 Societal Impacts of Predictive Analytics

The rise and usage surge of predictive analytics are the result of the digital revolution. Companies have invested in predictive analytics models to assist them with a variety of challenges with the belief that Big Data is the solution. Despite these advances, which assist in efficiency and automation, many people do not realise its societal impact. Billions of people are unaware of the prevalence of predictive analytics in their daily lives and have no idea if it is having a positive or negative effect.

5.1 How We Get Information

Prior to the digital revolution and the introduction of the home computer in the 1970's, people would get their information from school, work, and library or by word of mouth. The environment would shape their world and the way they see it. With the introduction of the computer and later, the creation of the worldwide web and hand-held devices, people began to have access to a whole new world with the touch of a button, anywhere and anytime.

Figure 6: *A graphic depicting the Internet*

Internet usage has increased exponentially and now services 56.1% (4.35 Billion) of the world's population (Miniwatts Marketing Group, 2019). This has has resulted in an overload of online information. Companies such as Google, Yahoo, Facebook and YouTube are using predictive analytics to help people navigate this endless amount of online information. Google has a market share of 89.95% of all online search engines, which makes it the most visited website in the world ahead of YouTube which is owned by Google's parent company, Alphabet (Statista). These Internet conglomerate have become the primary source for information and news for billions of people around the world. In the past, people researched information in a library or received their information from newspapers, radio or television. Most of us have become accustomed to clicking on Google or YouTube when searching for information. These Internet search engines use statistical algorithms to produce the best search results. These algorithms incorporate a variety of criteria for a search, but one of the main factors for Google search's is popularity of sources and the amount of traffic a source or website might have. In addition to popularity, Google uses previous search histories for future searches, which can create in an information loop (wsoaonline, 2015). This can result in the filtering out of websites/articles that are not used very often or those with opposing viewpoints, promoting a "group think" mentality. In many ways this is the goal of these websites. Giving people what they "think" they want, while excluding information, without many users control or knowledge.

As a result, we as a society have given these companies the authorization and responsibility of choosing what information they deem to be important. Since the goal of most companies is to make money, a company's interest can conflict with the best interest of there users. By allowing these companies to determine what we see or don't see, based on predictive analytics and data driven algorithms has the potential to be detrimental to the free flowing of ideas online.

5.2 Public Spending

Local government use predictive models to help improve efficiency when it comes to spending public funds. Governments at all levels are constrained by their budgets and are trying to stretch the taxpayers' money. When it comes to health care spending, infrastructure development and policing local government have been relying on predictive analytics models to optimize the decision-making process.

In terms of health care spending predictive analytics has been a great tool to help hospitals reduce weight times and improve outcomes (EHRIntelligence. 2019). Potentially, hospitals can gather a tremendous amount information which can help determine where to allocate resources, based on the needs of the populations they serve. This can be problematic because of confidentiality. It can be unethical to collected patient information without consent. If data is not collected appropriately, outcomes can be skewed and invalid. As we know from recent scandals with private companies such as Equifax and Target, which have experience major data breaches where information was stolen, we know that people are becoming increasingly more aware and concerned about how their information is being stored and used.

Predictive analytics in policing is being used more frequently in big American cities to help track and predict crime patterns. Public policies are being made, based on the findings of the statistical software. Areas with higher crime are starting to be allocated more police resources to curve the crime in those areas. Statistical models are only as good as the information that is inputted into them. In practice, police departments are trying to input information about all crimes, including petty crimes such as "vagrancy, aggressive panhandling, and selling and consuming small quantities of drugs. Many of these "nuisance" crimes would have gone unrecorded if a police officer was not there to see them" (ONeil, 2018, p.76). This can lead to the feedback loop of crime by criminalizing petty crimes and not focusing on the more serious crimes, such as murder, assault and arson, which are of greater concern, for the safety of a community. Petty crimes are more closely linked to the poverty-stricken neighbourhoods and an increased police presence would lead to a higher incarceration rates for these types of crimes (ONeil, 2018). There is a correlation between higher crime rates and in poverty-stricken neighbourhood. Well-off people tend to move away from these areas. The exodus of money from an area, can lead to a further increase in crime: lack of funding for a public school can be linked to less income from property taxes. This can cause an endless loop of crime and poverty in an area which needs help.

6 Conclusion

Predictive analytics has sparked a data driven revolution, changing the way companies, government and people are making decisions. The use predictive analytics models is becoming more prevalent and help government decision makers. Predictive analytics have expanded societies capabilities in terms of efficiency and has revolutionized the age of big data. If we look at the most successful companies in the world today, most of them are in the technological sector where big data and predictive analytics rule. Companies with an understanding on how to use predictive analytics, can help with optimal decision making and improve their profits to meet the needs of consumers.

Despite the tremendous capabilities of predictive analytics, it is important to take a step back and look at the societal impacts of predictive analytics and its ethical implications. Predictive analytics models are only tools which that help companies and organizations meet a certain objective. If those objectives are flawed or are biased against people's best interest there can be negative consequence, unbeknownst to the public. People and society do not realize that their options and choices are becoming more and more limited depending on how predictive analytics models are used. In reality society may not be controlling these models and models maybe controlling them.

7 Works Cited

7.1 Written Works

1. Abbot, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst.* Hoboken, NJ: Wiley.
2. Crawford, K., & Schultz, J. (2014). *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms.* Boston College Law Review,55(1).
3. A. O., Benjelloun, F., Lahcen, A. A., & Belfkih, S. (2017). *Big Data technologies: A survey.* Journal of King Saud University - Computer and Information Sciences,30(4). Retrieved March 22, 2019.
4. *Does Your Internet History Effect Google Search Results?* (2015, April 07). Retrieved March 30, 2019, from https://www.wsoaonline.com/does-your-internet-history-effect-google-search-results/
5. *Search engine market share worldwide.* (n.d.). Retrieved March 28, 2019, from https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/
6. *EHRIntelligence.* (2019, March 18). Enabling Targeted Health Data Exchange for Efficient Patient Care. Retrieved March 30, 2019, from https://ehrintelligence.com/news/enabling-

targeted-health-data-exchange-for-efficient-patient-care

7. Kelleher, J. D., Mac Namee, B., D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (1st ed.). Cambridge, MA: The MIT Press.
8. *History.* (n.d.). Retrieved from https://www.lloyds.com/about-lloyds/history
9. Eckerson, W. W. (2007). *Predictive Analytics: Extending the Value of Your Data Warehousing Investment.* TDWI Best Practices Report.
10. Nyce, C. (2007). *Predictive analytics white paper.* Malvern, PA: AICPCU.
11. Ioannidis, S., Montanari, A., Weinsberg, U., Bhagat, S., Fawaz, N., & Taft, N. (2014). *Privacy Tradeoff in Predictive Analytics.* Retrieved from https://arxiv.org/pdf/1403.8084.pdf.
12. *Punched Cards to Petaflops.* (2013). National Security Science,35-41. Retrieved from https://www.lanl.gov/discover/publications/national-security-science/2013-april/_assets/docs/punchcards-petaflops.pdf.
13. *Underwrite.* (n.d.) In Merriam-Webster's. Retrieved from www.merriam-webster.com
14. ONeil, C. (2018). *Weapons of math destruction: How big data increases inequality and threatens democracy.* London: Penguin Books.
15. *World Internet Users Statistics and 2019 World Population Stats.* (n.d.). Retrieved March 28, 2019, from https://www.internetworldstats.com/stats.html

7.2 Figures

Cover graphic: [Digital image]. (n.d.). Retrieved April 8, 2019, from https://www.tno.nl/en/focus-areas/information-communication-technology/expertise-groups/data-science/

Figure 1: [Digital image]. (n.d.). Retrieved from https://www.recruitmentgrapevine.com/content/article/2015-05-29-new-head-of-talent-sourcing-at-lloyds-of-london

Figure 2: [Digital image]. (n.d.). Retrieved from https://www.lynda.com/SPSS-tutorials/Essential-Elements-Predictive-Analytics-Data-Mining/578072-2.html

Figure 3: [Digital image]. (n.d.). Retrieved from https://www.electronicdesign.com/embedded-revolution/neural-network-hardware-drives-latest-machine-learning-craze

Figure 4: [Digital image]. (n.d.). Retrieved April 06, 2019, from https://www.computerhistory.org/revolution/analog-computers/3/143/311

Figure 5: Koblin, A. (n.d.). Flight Patterns [Digital image]. Retrieved April 6, 2019, from https://newatlas.com/art-ones-and-zeros-data-visualization/49926/#gallery

Figure 6: [Digital image]. (n.d.). Retrieved April 8, 2019, from https://www.cigionline.org/publications/one-internet

8 Appendix

Team 42 would like its team members to be marked individually. The breakdown of authorship between members of Team 42 is as follows:

Finn Plummer: executive summary, introduction, all subsections of section 3

Joshua Goldsmith: methodology, conclusion, all subsections of section 5

Brennen Creighton-Young: cover page, table of contents, appendix, works cited, all subsections of section 4, document formatting