4-21-2017

The Technological Revolution and Data Science

Leslie Walcott
Western Michigan University, 2013lwalcott@gmail.com

The Technological Revolution and Data Science

Leslie Walcott

Spring 2017

TABLE OF CONTENTS

Introduction and Purpose:

What was once only depicted in science fiction is now a reality: computers are taking jobs from humans. As technology improves, automation is transforming the workplace. They say a "fourth industrial revolution" is inevitable within the next ten years. In the industrial revolution, the jobs lost were unskilled laborers, such as coal miners, textiles manufacturers, or cotton workers. There was no argument for whether or not a machine could do the jobs more efficiently--it was fact. The term technological unemployment means the loss of jobs caused by technological change. The headline, "Factory workers replaced by automation," is not particularly startling to anyone. What may be surprising is what they say may happen in next next few decades: the replacement of highly skilled jobs with computers and robots.

Just this year, a law firm hired the first ever artificially intelligent attorney (De Jesus). Jill Watson is an A.I. teaching assistant who successfully assisted an entire semester at Georgia Technological University (Hill). A.I.'s have the capacity and potential to do more than we have ever dreamed. AI's are "taught" through a process called machine learning, which is a way of teaching a computer to learn from information which it already has.

There are some who think it is possible for data scientists to be among those who lose their jobs to computers and machines. The purpose of this paper is to present counter-arguments to this idea. Data science is a science driven by creativity, and we believe that there are human components to it which a computer cannot do. There are limitations of the computerization of machine learning and data analysis. This paper will delve into some of these limitations, and paint a picture of the evolution of the career. Creativity, intuition, and interpretation are vital skills of data scientists and driving forces behind why artificially intelligent beings will not take these jobs.

Evolution of Technology:

Around the 18th century, the Industrial Revolution transformed the world's working class. During this time, many important inventions were made and the transition to new manufacturing processes took place. In many industries, the work was previously done by hand. Oliver Evans invented an automated flour mill in the mid 1780's that used conveyors so that no labor was needed from the time grain was loaded into the elevator buckets until the flour was discharged into a wagon (PBS). Evans spent at least seven years perfecting five machines that formed a production line, replacing many workers in the industry. In 1790, Samuel Slater opened the first industrial mill was opened in the United States, which increased the speed with which cotton thread could be spun into yarn. This was all a part of the initial shift from handmade to machine made products. This is the time when there was an uprising in fears over the impact of machinery on jobs, as the unemployment rate grew. Some economists thought that overall innovation and technological advancements would have a positive effect on the overall job climate in the long term, but only time could prove this true or false.

As the 19th century began, worldwide industrialization continued. Evans, mentioned earlier, invented his lifetime's most important invention in 1803. His steam-powered engine could drive 12 saws through a hundred feet of marble in 12 hours. Evans's machine was light and portable, and could "transport 100 barrels of flour from Lancaster to Philadelphia in two days instead of three, tripling profits"(PBS). There was a widespread production of goods by machines. This resulted in reducing labor costs, thereby lowering prices. Many concerns over the negative impact of innovation disappeared. Classical economists formalized their arguments that technological advancements would not have a negative effect on jobs. It became increasingly apparent that technological progress was benefiting society, including the working class. The term Luddite fallacy was coined (Wikipedia contributors). The Luddites were a group of English textile workers who destroyed machines because they feared that these machines were going to take their jobs. The Luddite fallacy is the observation that new technology does *not* lead to higher overall unemployment in the economy--it just changes the job composition.

As the 20th century went on, many economists warned about technological unemployment. Yet a clear majority of both professional economists and the interested general public held the optimistic view through most of the 20th century. John Maynard Keynes wrote a broadly optimistic essay, "*Economic Possibilities for our Grandchildren*". It imagined a middle way between revolution and stagnation that would leave the said grandchildren a great deal richer than their grandparents. One of the worries Keynes admitted was a new "disease": "technological unemployment…due to our discovery of means of economising the use of labour outrunning the pace at which we can find new uses for labour."

Today, computers and robots taking over jobs is discussed more than ever. You can walk into a fast food restaurant and have a computer take your order, rather than a cashier. In 2016, McDonald's rolled out their "Create Your Taste" touch screen ordering kiosks. Panera Bread installed similar kiosks into 50% of their locations, and expect to have them in all stores within a year (Johnson). Many restaurants have announced plans to follow suit, including Wendy's and Hardee's. An in-depth study by Oxford University and Deloitte examined "how susceptible jobs are to computerisation", using a program that analyzed data from 702 occupations. A selected portion has been included in Figure 1.

Probability	Occupation
0.0031	Mental Health and Substance Abuse Social Workers
0.0035	Occupational Therapists
0.0042	Physicians and Surgeons
0.0055	Human Resources Managers
0.0065	Computer Systems Analysts
0.0077	Anthropologists and Archeologists
0.009	Registered Nurses
0.0095	Teachers and Instructors, All Other
0.014	Engineers, All Other
0.015	Computer and Information Research Scientists
0.015	Music Directors and Composers
0.021	Photographers
0.035	Lawyers
0.039	Political Scientists
0.082	Graphic Designers
0.13	Software Developers, Systems Software
0.15	Electricians
0.18	Public Relations Specialists
0.22	Computer Occupations, All Other
0.22	Statisticians
0.35	Flight Attendants
0.35	Plumbers, Pipefitters, and Steamfitters
0.37	Actors
0.43	Economists
0.44	Historians
0.51	Dental Assistants
0.54	Massage Therapists
0.61	Market Research Analysts and Marketing Specialists
0.65	Librarians
0.67	Bus Drivers, Transit and Intercity
0.68	Postal Service Mail Carriers
0.72	Home Appliance Repairers
0.77	Bartenders
0.77	Dishwashers
0.80	Barbers
0.81	Word Processors and Typists
0.83	Railroad Conductors and Yardmasters
0.84	Tool and Die Makers
0.85	Sales Representatives, Wholesale and Manufacturing, Except Technical and Scientific Products
0.86	Real Estate Sales Agents
0.91	Musical Instrument Repairers and Tuners
0.92	Insurance Sales Agents
0.92	Retail Salespersons
0.94	Accountants and Auditors
0.94	Paralegals and Legal Assistants

0.95 Bill and Account Collectors
0.97 Cashiers
0.98 Loan Officers
0.99 Watch Repairers
0.99 Telemarketers

Figure 1

Evolution of Data Science:

As technology improved and the computerization of jobs became more and more common, a new career emerged. The job title data scientist has not been around nearly as long as mathematician or statistician. The term was coined relatively recently with the advancement of computer science and machine learning. In 1962, mathematician John Tukey published "*The Future of Data Analysis*" (Press). Perhaps the most important part of the paper was the development of the idea of exploratory data analysis (EDA). EDA is the basis of data analysis, a new approach to analyzing data sets that relies more on intuitiveness and visualization than strict mathematical equations. Prior to the publication, scientists in the field focused on initial data analysis (IDA), which places the primary focus on checking the statistical and mathematical assumptions required for hypothesis testing. For this classical method (IDA), the process looks like this: Problem => Data => Model => Analysis => Conclusions. For EDA, the sequence is: Problem => Data => Analysis => Model => Conclusions (Croarkin). With classical IDA methodology, a model is immediately chosen, and the analysis and testing that follows are focused within the already known parameters of that model. EDA changed that, and data collection was followed by analysis and visualization techniques in order to choose the most appropriate model. Tukey's goal was to encourage scientists to explore the data. His paper closes with, "What of the future? The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it?...Who is for the challenge?"

Sweden's Peter Naur was another great contributor to the field. He disliked the term "computer science", and was the first to call the emerging field "data science" (Augur). He used the term in his book, *Concise Survey of Computer Methods*, published in Sweden and the United States in 1974. His own definition of the term was, "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences." A lot happened in the world of data science in the 1970's: in 1977, the International Association for Statistical Computing was founded. At the time, their mission statement was to "link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge". It has since been updated to include goals of "fostering world-wide interest in effective statistical computing."

Today, data science is frequently used alongside the term "big data". This term was coined around the time that most of the public gained access to the Internet (the very end of the 20th century). An article published in 1988 addressed this idea of monumental data growth, saying "the growth rate of traffic on the public Internet, while lower than is often cited, is still about 100% per year, much higher than for traffic on other networks," (Press). 1999 may have seen the first published use of the term "big data", with the publication of "*Visually exploring gigabyte data sets in real time*", by Steve Bryson (Press). The article opens with the following statement: "Very powerful computers are a blessing to many fields of inquiry. They are also a curse; fast computations spew out massive amounts of data. Where megabyte data sets were once considered large, we now find data sets from individual simulations in the 300GB range. But understanding the data resulting from high-end computations is a significant endeavor. As more than one scientist has put it, it is just plain difficult to look at all the numbers. And as Richard W. Hamming, mathematician and pioneer computer scientist, pointed out, the purpose of computing is insight, not numbers." This certainly melded well with Tukey's ideas. The meaning of "big data" as we use it today didn't come about until much later, around 2007 or 2008, and today the average US company with over 1,000 employees is storing more than 200 terabytes of data according to the report Big Data: The Next Frontier for Innovation, Competition and Productivity by McKinsey Global Institute (Herold).

As mathematics, statistics, and computer science combined into data science, a new career emerged: data scientist. Data science has been increasingly adopted by companies in all fields. Every field produces data, and therefore can benefit from improving the way they analyze their data. There are the obvious fields: digital marketing and targeted advertisements. Digital ads are targeted based on a user's search and browsing history, all the result of data science algorithms. Netflix and Amazon are able to recommend what you might want to watch or buy next, due to their analysis of past trends. The healthcare field is also quickly adopting data science (Henke). In 2012, the National Institute of Health awarded $15 million in funding for 8 projects to research uses of big data in the healthcare world. Biostatisticians substantiate claims by conducting and analyzing studies and trials. Product companies are outsourcing their data to companies like Cloudera, searching for ways to use their massive data sets in a better way.

In 2015, 43% of companies surveyed reported a lack of appropriate analytical skill in their employees. The McKinsey Global Institute produced a report in late 2016 titled "The Age of Analytics: Competing in a Data-Driven World". The report says degrees granted in data science fields has grown by 7.5 percent between 2010 and 2015. The demand in the job field for data scientists is growing perhaps as much as 12 percent per year. The net effect, they conclude, is that the U.S. economy could be short as many as 250,000 data scientists by 2024.

The Technological Revolution:

As discussed earlier, the technological revolution is the replacement of one job (or technology) with a different, improved technology. It has clearly had a drastic effect on many industries, with more jobs becoming computerized each year. Could machine learning render data scientists obsolete? Computers have replaced cashiers, a teaching assistant, a lawyer--what is stopping a computer from soon replacing a data scientist? A data scientist's job is much more than simply running a computer program and reading an output. This is part of what makes the job so unique. It is not only about an algorithm, plugging in numbers, and reading a linear model. If the job was simply taking data, plugging them in, and copying an output, then the risk of computerization would be much, much higher--it would have already happened. There are many human aspects of the job that are absolutely vital. The career itself is ever-evolving, and requires deep creativity and intuition that cannot be replicated by a machine. The next sections will delve into some of the necessary human aspects of the career: communication, collaboration, and context. Each of these have roots in creativity, and help to bridge the gap between science and art. Data science is about more than the scientific tools (linear regression, decision trees, deep learning). The scientists need to understand how to assemble all of these tools into something meaningful, and how to symbolize and communicate convincingly that it is meaningful to others.

Communication:

A very important skill of a data scientist is the effective communication of results and their implication. Bachelor and Master's programs in statistics and data science have courses devoted to this. As a data scientist, when you have results in hand, communicating those results effectively is perhaps the most important part of your job. Chances are, the people those results are the most important to are not going to be formally trained in your field, and therefore you must "translate" your results into language that everyone can understand the significance and implication of. Western Michigan University offers such a course, entitled "Communicating Statistical Results". The course description says, "The emphasis of the class will be the reporting of statistical analysis so that all relevant information is conveyed, avoiding the use of jargon and enhancing the text with the use of informative tables or graphics. Students will be assigned projects involving data gathering and analysis. Written and oral reports on the methodology used and the results of the analysis will be required of each student."

Communication is vital in workplace. Dr. Janice Derr, author of *"Statistical Consulting: A Guide to Effective Communication*", says that one of the biggest challenges faced by a data scientist is communicating information to those with less formal knowledge of the field. This brings up an important point: a data scientist must be much more than a skilled mathematician. The way they present their findings is nearly as important as the methods they employ to find them in the first place. People are overwhelmed with data--the last thing they want is a printed output with a bunch of numbers they need to make meaningful. People are so constantly bombarded with new data, facts, and figures that they begin to blend together into a constant background hum. Recognizing the issues with data overload can help set apart a study, presentation, or lecture from the rest. The goal of presenting your results is to make them memorable and repeatable, and this process of taking results from numerical to meaningful is inherently human. This is where the data scientists bridge science and art: when they take a scientific method, with a deep understanding of the derivations of the mathematics, and communicating those results in a way that draw people in. They have to be creative in the ways they present their information: data without the story and context behind it is just numbers. Data scientists have to take numbers and transform them into something that brings them to life, inspires people to action, and makes them care.

Collaboration:

Data science has quickly and fascinatingly become a collaborative career. When you have a data set consisting of 5,000,000+ lines, there is no one “best” model. There are thousands of models to be created, and one person may think one is better while another doesn’t. There are so many possible interactions and patterns to be found. There are an infinite number of approaches that apply to any problem. There is a lot of noise and richness to datasets, and working as a team can quickly uncover hidden relationships.

An interesting example of this is Kaggle, a crowd sourcing data modeling company. Kaggle began as a place for data sets to be posted, and data scientists can post their ideas, work, code, and models to build off of one another. The most infamous Kaggle dataset is about the sinking of the Titanic. They provide you with the dataset: a list of all the passengers on board, with information about each, including survival (yes/no), gender, ticket class, age, number of family members aboard, port of embarkation, and more. “Kagglers” use machine learning tools to predict which passengers survived the tragedy. This particular data set has been available for 4 years, and 5,781 teams have worked with it. Kaggle also hosts competitions, and large companies have submitted their own problems in order to have their data worked on by the most talented minds in data science. Awards for the winning model range from a cash prize to a job at Facebook. One of the most popular projects was done for Microsoft on their Kinect platform.

Kaggle players came up with algorithm which can be shown a gesture just once, and can then recognize it again later, for dozens of gestures (Howard). The “dataset” for this competition is actually a collection of sophisticated video clips taken by Kinect’s motion-detecting camera. The winning team won $10,000. Companies benefit from using this crowdsourcing approach to data science: they get the best result at a low cost. The “best” results come from the teamwork: the community of experts use Kaggle as an “office water cooler” to collaborate and exchange ideas. This collaboration leads to realizations and inferences that otherwise may not be made. If one person on a team or as a competitor makes a discovery important to the dataset, others will follow suit.

Ed Roth, an Associate Professor of Music Therapy at Western Michigan University, has given some interesting talks about the role of intuition and collaboration in both art and science. There are a number of ways we collaborate with others: musical improvisation as a group, a sports team creating plays on the fly, mathematicians working through a difficult proof. These collaborative efforts produce oxytocin in the brain, a neurotransmitter that “creates both the chemistry and physical structure in the brain to generate and be open to new ideas”. Some neuroscientists refer to this as “the Flow”. Great collaboration leads to great ideas, and encourages scientists to test their own theories with the risk that it’s okay to explore.

Florian Douetteau, cofounder and CEO of data science start-up Dataiku, says "Think of real-time collaboration on designing a data flow, where people can build a part of the flow, connect their parts together with other parts, and assign tasks to others on the team," he explains.

"...Collaboration is important for teamwork just because you need to be able to see what other people are doing. When you're able to see what someone is doing *while* they're doing it, you get ideas. You can suggest a better way of doing it; you can point out other [factors] they should consider.” Linda Hill, professor of business at Harvard University, tells us, “Innovation is not about solo genius, it's about collective genius." Collaborative efforts produce a creative and open environment, the perfect breeding ground for new ideas and success.

Context:

Something very important to consider while analyzing data is context. This is something that would be difficult for computers to replicate. If you give a computer a data set including the location of fires and firefighters in a certain city, the output may say that firefighters cause fires. Of course, we know this is ridiculous, but certain contextual clues are not as obvious, and this is why data science isn't as simple as plugging in numbers and copying the model output. Jeff Jonas, engineer and chief scientist for IBM Entity Analytics, is outspoken about this issue, giving talks nationwide about the vital role of context in data science. He has said, "Drawing insight from a piece of data involves understanding how it fits into the larger picture of an organization." The key of data science is not the parameter estimates in the output of a model, but rather, how that model affects a company's day to day operations, and what will change as a result of your findings.

Jonas runs a blog where he discusses some of his work, methods, and thoughts on the growing industry. He has written about the importance of context on multiple occasions, saying, "Operating on a datum without first placing it into context is a risky proposition. And thus, from my point of view, determining context is the most significant technical hurdle necessary to deliver the next generation of business intelligence." Data with absolutely no context is, of course, worthless, to both man and computer. Data with limited context can give you a model that is either not quite right or just plain wrong. The more data you have to draw information from, the better. Imagine a company that analyzes their internal data involving their sales. While they may be able to build models and analyze their growth and other figures, they aren't using data science to its full potential with the limited context of their data (Lorentz). Their internal data is isolated from the rest of the data in the world, involving major news stories, weather patterns, historical information, social media, or even an election. Rather than concluding their sales fell last month and they should push their product harder, they may conclude their slipping sales had to do with recent elections and uncertainty, and make a plan for their business for the next time there is a shift in the political climate.

The ability to see the big picture is vital in getting the most out of your data science. The goal is to analyze data and "detect human-based meaning from it". This goes back to the data scientist's job of communicating their results. They have to take into consideration the context surrounding the problem before deciding how to best proceed with the problem, and then share their findings with others.

Case Study: Election of 2016

In 2016, people were paying attention to polling and data far more than usual: it was an election year. As November drew near, the general consensus was that the new president of the United States would be Hillary Clinton. Virtually every major vote forecaster had had Clinton winning. Nate Silver's popular FiveThreeEight site gave Trump the best chance of winning at a meer 29% (Vogel). Other models, including the New York Times Upshot and the Princeton Election Consortium, had Trumps odds at 15%, 8%, 2%, and even less than 1% (Lohr). Election betting odds had Trump's chances at only 18% at midnight Tuesday. People have placed a lot of faith in data, and it was a tough night for those crunching numbers. How did everyone get it so wrong?

Predictive analysis remains a relatively young science. The United States only elects a president every 4 years, and the world drastically changes in the time, rendering past models and strategies useless. Pollsters can still ask people about their intentions to vote, but even that has changed: just recently, this was all done via landline. This year, a lot more polling was done online, bringing another issue into play: how to be sure that your responses are coming from real people and not a bot. There is also, of course, no way to poll every potential voter, so you hope your samples are as representative of the population as possible. But how do you ensure this? People who are still have landlines and are polled this way are much more likely to be from the older generation, and those who are active enough on the internet to get digital polls are likely younger, or work in certain professions. Dan Zigmond, a data scientist at Facebook, was asked about his thoughts of the incorrect models leading up to the election. "As a data scientist, I always think more data is better. But we really don't know how to interpret this data," Zigmond said. "It's hard to figure out how all these variables are related." An article on Wired articulated a particularly important point: "A deep neural network can't forecast an election unless you give it the data to make the forecast, and the way things work now, this data must be carefully labeled by humans for the machines to understand what they're ingesting." There are so many confounding, human factors that must be analyzed before a computer can spit out a model.

This election was unique in many ways. With perhaps the most controversial candidate ever, a new problem was introduced: people not being honest in the polls. The Politico Caucus is a group of activists and strategists in 11 battleground state, and 71% of Republican insiders said that polls understated Trump's support because voters "didn't want to admit to pollsters that they were backing the controversial candidate" (Shepard). There are countless reasons why an individual may not vote for Trump, due to his statements and actions against many different groups of people. There were also a large number of Republicans who pledged to never support Trump during the primaries, never thinking he would be the candidate in the general election.

There are countless factors that cannot be labeled, plugged into an Excel worksheet, and then imported into a statistical programming suite in order to get a result. Data and the context surrounding it is far too complex for that. A number of these have been reported and commented on since the election. On October 28th, 2016, FBI director James Comey announced the discovery of new emails that seemed important to his closed investigation of Clinton. On November 6th, two days before the election, Comey absolved Clinton of wrongdoing--after millions of votes had already been cast early. Some states did not polls after October 28th. Failing to poll after a decisive event had taken place surely resulted in some numbers being inaccurate. States that showed some of the biggest surprises also tended to be states where there were fewer polls toward the end of the race (McElwee). Comey's letter could be one of the critical factors in the election. This is just one example of factors that go into data science that aren't as simple as inputting numbers and variables.

Computerization of data science has been proven to do some simple things - improve the odds of making a sale through targeted ads, for example. But using data science in more industries with higher stakes require humans to deeply analyze, study, and understanding much more than just numbers. There are assumptions and decisions that are made based on infinite factors that all affect the outcome. The model building done by a computer is just one of many tools used by data scientists and analysts.

Conclusion:

In a time where technology is changing faster than ever, it is impossible to predict what will happen next. Machine learning has dramatically changed the computerization landscape, and more changes lie ahead. Despite this uncertainty, we believe that data scientists can rest knowing their job is safe. The career of a data scientist has evolved to include deep mathematical knowledge, an understanding of the business landscape, intuition, and above all, creativity. The career encompasses both science and art. Anthony Goldbloom, founder and CEO of Kaggle, gave a Ted Talk entitled "*The jobs we'll lose to machines--and the ones we won't.*" He leaves us with a quote, inspiring to all those thinking of their future: “Whatever you decide to do, let every day bring you a new challenge. If it does, then you will stay ahead of the machines.”

Works Cited

Augur, Hannah, Alex Bordei, Eileen McNulty, and Manu Jeevan. "Beginner's Guide to the History of Data Science." *Dataconomy*. N.p., 30 May 2016. Web. 18 Feb. 2017.

Croarkin. "1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?" *1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?* N.p., n.d. Web. 10 Mar. 2017.

De Jesus, Cecille. "AI Lawyer "Ross" Has Been Hired By Its First Official Law Firm." *Futurism*. N.p., 05 Sept. 2016. Web. 10 Feb. 2017.

Frey, Benedikt, and Osborne. "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change* 114 (2017): 254-80. Web. 2017.

Goldbloom, Anthony. "Transcript of "The jobs we'll lose to machines -- and the ones we won't"." *Anthony Goldbloom: The jobs we'll lose to machines -- and the ones we won't | TED Talk Subtitles and Transcript | TED.com*. N.p., 2016. Web. 10 Aug. 2017.

Henke, Bughin, Chui, Manyika,Saleh, Wiseman, and Sethupathy. "The age of analytics: Competing in a data-driven world." *McKinsey Global Institute & Company*. N.p., Dec. 2016. Web. 20 Mar. 2017.

Herold, Benjamin. "In Analytics Economy, Demand for Data Scientists Outpaces Supply, McKinsey Says." *Education Week - Digital Education*. N.p., 08 Dec. 2016. Web. 20 Apr. 2017.

Hill, David J. "AI Teaching Assistant Helped Students Online-and No One Knew the Difference." *Singularity Hub*. N.p., 03 June 2016. Web. 10 Feb. 2017.

Howard, Jeremy. "Only Kinect: More than a Gaming Gadget." *No Free Hunch*. N.p., 17 May 2016. Web. 9 Mar. 2017.

Internal Association for Statistical Computing. "About IASC." *IASC*. N.p., n.d. Web. 10 Mar. 2017.

Johnson, Hollis. "Fast food workers are becoming obsolete." *Business Insider*. N.p., 16 May 2016. Web. 1 Mar. 2017.

Jonas, Jeff. "Context: A Must-Have and Thoughts on Getting Some ..." *Jeff Jonas*. N.p., 1 July 2007. Web. 6 Feb. 2017.

Keynes, John Maynard. *Economic possibilities for our grandchildren*. Seattle, WA: Entropy Conservationists, 1991. Print.

Lohr and Singer. "How Data Failed Us in Calling an Election." *The New York Times*. The New York Times, 10 Nov. 2016. Web. 19 Mar. 2017.

Lorentz, Alissa. "With Big Data, Context is a Big Issue." *Wired*. Conde Nast, 06 Aug. 2015. Web. 10 Mar. 2017.

Marr, Bernard. "A brief history of big data everyone should read." *We Forum*. N.p., 2015. Web. 9 Jan. 2017.

Metz, Cade. "Trump’s Win Isn’t the Death of Data-It Was Flawed All Along." *Wired*. Conde Nast, 09 Nov. 2016. Web. 1 Apr. 2017.

New York University. "What is data science." *Data Science at NYU*. N.p., n.d. Web. 10 Mar. 2017.

O'Neil, Cathy. *Weapons of Math Destruction*. S.l.: Penguin , 2017. Print.

Press, Gil. "A Very Short History Of Data Science." *Forbes*. Forbes Magazine, 15 Oct. 2014. Web. 20 Mar. 2017.

Press, Gil. "The Supply And Demand Of Data Scientists: What The Surveys Say." *Forbes*. Forbes Magazine, 16 May 2015. Web. 19 Mar. 2017.

McElwee, McDermott, and Jordan. "4 pieces of evidence showing FBI Director James Comey cost Clinton the election." *Vox*. Vox, 11 Jan. 2017. Web. 30 Mar. 2017.

Shepard, Steven. "GOP insiders: Polls don't capture secret Trump vote." *Politico*. N.p., 28 Oct. 2016. Web. 10 Mar. 2017.

Siegel, Ethan. "The Science Of Error: How Polling Botched The 2016 Election." *Forbes*. Forbes Magazine, 09 Nov. 2016. Web. 28 Mar. 2017.

Silver, Nate. "Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else." *FiveThirtyEight*. FiveThirtyEight, 11 Nov. 2016. Web. 28 Mar. 2017.

Swoyer, Stephen. "Is Collaboration the Critical Success Factor for Data Science?" *TDWI Upside*. N.p., 25 Aug. 2016. Web. 10 Mar. 2017.

“Technological Revolution.”*Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc. 5 January 2017. Web. 10 Jan. 2017.

"The onrushing wave." *The Economist*. The Economist Newspaper, 18 Jan. 2014. Web. 10 Feb. 2017.

Thompson, Derek. "The Next Industrial Revolution." *The Atlantic*. N.p., 6 Sept. 2016. Web. 19 Mar. 2017.

Tukey, J. W. (1962). The future of data analysis. Annals of Mathematical Statistics, 33, 1–67.

Tukey, J. W. (1977). Exploratory data analysis. Reading, PA: Addison-Wesley. Velleman, P. F., & Hoaglin, D. C. (1981)

Veazey, Simon. "The Impact of Technology's Invisible Hand." *The Epoch Times*. N.p., 31 Oct. 2016. Web. 10 Feb. 2017.

Vogel and Isenstadt. "How did everyone get it so wrong?" *Politico*. N.p., 9 Nov. 2016. Web. 10 Mar. 2017.

"Who Made America?" *PBS*. Public Broadcasting Service, Sept. 2014. Web. 10 Apr. 2017.