CS 445 Week 6 - Research Paper - Big Data

Matthew T. Hartigan

Regis University

‘Big Data’ is one of the most popular and controversial topics of conversation in technology in the modern world. Since the advent of computers in the 20th century, mankind has been steadily improving the hardware and software that we use to generate and manage increasingly large sets of data. The main frame has given way to the personal computer, which has given way to the laptop, and now society has reached the point where it is difficult to find someone who does not have a mini-computer resting in their pocket in the form of a smartphone. All these devices make it easier than ever to generate, collect, and transmit data about virtually every industry. These data sets have now grown so massive that they cannot be handled by traditional database management systems and processes. These new, massive data sets are what we call ‘Big Data’. In this essay, we will explore further the topics of exactly what Big Data is, why we need it, what dangers are associated with it, what modern tools are used to work with it, and who is using it around the globe.

Oracle, a major technology company that is intimately involved with data management, defines Big Data as data that ‘contains greater variety, arriving in increasing volumes and with ever-higher velocity” (None, 2019). Human beings have been collecting data for thousands of years. At first data collection was as simple as writing notes on a stone tablet or keeping a notebook of store inventory. When computers were invented in the 20th century, this data collection began to kick into high gear because traditional, hand-written records could now be stored digitally. In the early 2000’s, content aggregation and social media sites such as YouTube and Facebook further accelerated the variety, volume, and velocity of data we collect daily. Today, people and companies are connecting previously inanimate objects to the internet. These objects (cars, fitness trackers, watches, etc.) generate even more data, and make up what we call the ever-expanding internet-of-things (Burgess, 2018). The result of this historical progression are the huge data sets that make up the Big Data phenomenon that Oracle described.

People and companies need Big Data because it helps drive better decision making, which leads to increased productivity and profits. As we explored in the preceding paragraph, it is no longer difficult to collect information regarding anything from your personal health to the emissions from your car. However, a massive data set all by itself is not very useful. Data becomes useful when we can use it to analyze trends, identify patterns, and ultimately learn things about an object, person, or concept that were previously invisible to us. One example of how Big Data has been successfully implemented to support better decision making is the oil and gas industry. Major oil producers like Shell own and operate staggering amounts of machinery in diverse, challenging environments, all around the world. Since Shell’s bottom line is tied to how efficiently and consistently they can extract oil and gas from the earth and bring it to market, unreliable or broken machinery is of major concern to them. Before Big Data, Shell had to manage their machinery in a ‘reactive’ manner. It was standard practice to keep spare parts on hand to fix a machine when it went down, only to find that those parts had gone missing, or were unusable due to rust when the time came. Once Shell was able to partner with companies that understood Big Data and how to connect their machinery to the internet of things, they began to turn things around. Instead of reacting to broken machinery and scrambling employees to fix it post-haste, Shell began to employ office-based workers to analyze their machinery data before it was needed. These analyses provided insights that allowed Shell to optimize their inventory management, identify trends that contributed to machinery breakdowns, root out poorer-performing devices, and much more. All these improved business practices were made possible only by Big Data and the hidden relationships that it enabled Shell to visualize (Boulton, 2018).

In *The Human Face of Big Data*, Richard Smolan presents several examples of how Big Data has been used to improve humanity; however, he also cites several instances where Big Data has been used in a negative way. One such danger that comes with using Big Data is data privacy. One famous example of how Big Data can be dangerous is the 2018 Cambridge Analytica scandal. Cambridge Analytica was a company that designed a Facebook survey that contained questions that, when answered, gave them insight on the preferences and tendencies of different individuals. Because survey takers were required to login to the survey using their Facebook account, Cambridge Analytica was also allegedly able to scrape personal information about these users and their friends, resulting in a massive collection of information about everyone’s likes, dislikes, and overall tendencies. Combining these data sets gave Cambridge Analytica the power to identify valuable patterns that they claimed they could use as input to how to design targeted ad-campaigns that could more effectively shift the actions of consumers and voters. The story caught headlines because it was allegedly used to support the 2016 Donald Trump presidential campaign. Regardless of which parties are telling the truth in this situation, and which may be stretching it, the fact remains that it is technically possible for individuals who are not associated with massive data aggregating companies such as Facebook, Instagram, and others to harvest data from social media sites and use it to gain massive influence on targeted audiences. There are countless examples of when this works out for the better. We all enjoy the recommended products section of Amazon. Auto-play and targeted ads on YouTube make video watching more fun. However, several ethical questions remain with regards to Big Data. Who owns all this data? Is it Facebook? Even if that is the case, are Facebook employees being screened with criteria to determine who is ‘safe’ to handle this kind of data? Furthermore, should selling data to third parties such as Cambridge Analytica be permitted? Clearly, there are many potential moral and ethical issues that can arise when it comes to Big Data. What’s even more challenging is that consumers / users often have little to no say in the matter, as was the case in the Cambridge Analytica story.

There are many tools currently available that are used to help manipulate Big Data. One of the first was the R language. Originally developed by researchers in New Zealand for statistical analysis, R has become one of the most widely programming languages when it comes to analyzing data and its associated trends. Statistical Analysis Software (SAS) was developed at the University of North Carolina for similar purposes. Today, high level languages such as Python and even Microsoft Excel are all used to help different groups of people work with the data they create. As the importance of Big Data continues to grow, new tools continue to be developed to meet the demand.

It is evident that Big Data is on an upward trajectory. As stated by Oracle, the volume of data is increasing, as it comes in at a higher velocity and with greater variety. Who is using all this data? Big Data’s influence stretches across industries. Whether it is a political campaign, an Oil & Gas company, or even a private individual who is simply concerned about tracking their own health outcomes, Big Data does not discriminate when it comes to who it is useful for. The phenomenon does not come without risk. As evidenced by the Cambridge Analytica saga, Big Data continues to pose challenging moral and ethical dilemmas to society, as we struggle to not only regulate what can and can’t be done with data, but we also struggle to anticipate which ethically compromising situations will crop up next before they happen. Although there are many problems with Big Data, there is no stopping it. From the humble beginnings of R and SAS, more tools such as Python packages, Hadoop, and others continue to be developed and rolled out to market in a seemingly never-ending attempt to satiate the world’s hunger for more information. In short, there is no stopping the Big Data train, we can hold on and try and steer it as we go along for the ride!

**References**

None. (2019). What is Big Data?.

Retrieved from: <https://www.oracle.com/big-data/guide/what-is-big-data.html>

Burgess, M. (2018, February 16). What is the Internet of Things? WIRED explains.

Retrieved from: <https://www.wired.co.uk/article/internet-of-things-what-is-explained-iot>

Boulton, C. (2018, November 14). 10 data analytics success stories: An inside look.

Retrieved from: <https://www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html>

None. [Digital Vidya]. (2018, February 7). Evolution of Data Analytics, Data Science & Big Data

Technologies. Retrieved from: <https://www.youtube.com/watch?v=Fe_OmzuNO2Q>