PAPER 1 – BDAA FALL 2016

Carlos Sathler
Graduate Student, MS in Data Science
Indiana University at Bloomington
cssathler@gmail.com

ABSTRACT

In this paper we answer a few questions about Big Data and contrast personal interests on the subject of Big Data among students attending the Fall 2016 on-line session of the Big Data Applications & Analytics class at Indiana University at Bloomington.

1. WHAT IS BIG DATA

Big Data refers primarily to the unprecedented vast amounts of data that our society has been exponentially accumulating particularly in the past 10 years, and will most likely continue to accumulate in the years to come. A now famous special report from the February 2010 issue of The Economist [1] speaks of a "data deluge" and contrasts an estimated 150 Exabyte of data created by mankind in 2005 with 1,200 exabytes of data estimated to be created 5 years later, in 2010; that is a 700% percent increase in just 5 years. In the 2016 Internet Trends Report Mary Meeker [2] compares the total amount of data in the "digital universe" from 2010 through 2016. She reports a compound annual growth rate in excess of 50% since 2010 resulting in an increase from a little over 1MM exabytes in 2010 to a whopping 9MM exabytes of data in 2016.

Understanding what technological changes propitiated this increasing capacity our society has developed to generate and accumulate data is useful in understanding Big Data and how it is different from “small data”. In the 2014 Internet Trends Report Meeker [3] highlights the decline of bandwidth, storage and computing costs, along with the rise of cloud computing. Together, these factors contributed to an explosion of availability and adoption of smart phones, tablets and a multitude of sensor enabled devices as diverse as wearable fitness trackers, smart meters, cars and drones, to name a few. Applications benefiting from these technologies and the high availability, high storage capacity, scalability and distributed computing inherent to cloud computing, explored the potential for crowdsourcing, various forms of monitoring, streaming, social networking, video and photo sharing, for example.

The data generated and processed by these applications is not only much bigger than “small data”; it has several other distinctive characteristics. Davenport [4] summarizes the most striking ones: “Big data refers to data that is too big to fit on a single server, too unstructured to fit into a row-and-column database, or too continuously flowing to fit into a static data warehouse” [p 1]. Another characteristic of Big Data lies in its potential to unleash new insights and discoveries. While in the past decision makers and scientists would formulate a question or hypothesis, and engage an analyst to look in the data for confirmation or denial of the hypothesis, with Big Data there is a shift to that approach; decision makers and scientists in possession of Big Data can now engaging a new type of professional, the data scientist, and ask “here’s the data, find out what it says about my problem”. This approach is sometimes called the “fourth paradigm” in science [5].

The type of applications that enables business leaders, scientists and the general public to draw insights and learn from large volumes of data is called “analytics” [6]. Analytics applications use inferential statistics and other techniques to correlate data in special ways in order to extract useful, actionable information from it.

Some applications of Big Data and analytics with tremendous impact to our lives include: news feeds on Facebook, Google search criteria auto corrections, personalized, targeted advertisement on the web, Netflix recommendations, fraud detection in banking systems and insurance, image recognition applications, customer sentiment analysis for helpdesk service improvement, automatic text translation, email spam filtering, disease prevention and diagnostics in healthcare, and many others.

Big Data and analytics offer the promise of increased profits, efficiencies and social good. So much so that business leaders and scientist currently see data as one of the most important assets to their businesses and organizations. In the article Data, data everywhere [7] Craig Mundie, now Senior Advisor to the CEO at Microsoft, spoke of a “data-centered economy” and stated that “data are becoming the new raw material of business: an economic input almost on par with capital and labor” [p 1].

2. WHY IS BIG DATA INTERESTING TO ME?

I have always nurtured a passion for data. In 1994 I came to the United States (from Brazil) to attend graduate school at Louisiana State University (LSU). One of the reasons I chose LSU was because Peter Chen, the inventor or ER Modeling, was a faculty at the Department of Computer Science at LSU. I had the privilege of attending one of his classes. Data modeling, systems thinking and systems analysis were my main interests when I started my career in IT more than 20 years ago, after attaining my engineering degree.

Around 2012 when I first started reading about Big Data I was naturally drawn to learn more about the changes in the way data was being created and used. From the start, Big Data meant to me the promise of human advancement through insights and discoveries in many areas of our lives that matter to me, such as health sciences, climate science, and social networking. I realize now that the main reason why I'm interested in Big Data is because of its potential to improve our lives in these areas.

Additionally, irrespective of applications, I'm very curious about the vast amounts of data that our civilization is amassing and what we will do with it. It is a curiosity about the data itself, and what it will "tell" us about the world we live in.

As I read through some of my colleagues' answers to the question above, I was able to identify a few students with similar interests. Many are interested in health-care data and applications; some are interested in science and education projects, or work in government; at least one is interested in supporting non-profit philanthropic organizations. Like me, many students also reported a curiosity about the data itself, irrespective of possible applications.

Finally, I have a third very important interest related to Big Data. I have been a project manager for about 15 years, managing customized off-the-shelf (COTS) ERP implementations, and also software development projects. I have always had a strong interest in methodologies in general and in particular software engineering methodologies. I may look for opportunities to work as a project manager on Data Science initiatives involving Big Data. For that reason I'm very interested in Data Science processes in the context of Big Data projects. I have not identified another student who has communicated that interest in Piazza.

3. WHY IS BIG DATA UNINTERESTING TO ME?

Several of my colleagues expressed a strong interest in learning how to use the tools that unleash the potential of Big Data. I have been a programmer for several years before becoming a consultant and project manager and I wouldn't like to again work as a programmer. However, I'm interested in developing a deep understanding of the technical landscape of the Big Data and analytics toolset, and I'm hoping to learn how to move data and process large datasets in the cloud; I'm also looking forward to writing at least one machine learning application using Python before I graduate, especially considering I had the opportunity to do so before using R.

4. WHAT LIMITATIONS DOES BIG DATA ANALYTICS HAVE?

Limitations of Big Data analytics include "pigeonholing" and lack of interpretability. Another significant limitation relates to Big Data unintended consequences. We will briefly discuss each of these limitations.

By "pigeonholing" I'm referring to a tendency that recommender systems have to offer too restrictive recommendations based on their interpretation of individual preferences. Probably every Spotify user has dealt with some level of frustration exploring playlists that are labeled "discovery" playlists but contain no meaningful variations over previously selected songs.

It is a known fact that Google personalizes search results [8] but how well can we expect Google to know how many conservative news sources a liberal person will want to see at any given time they look for political news update? More generally, how can a recommender system suggest something meaningfully different, yet only meaningfully different in a way that will still satisfy the user?

On the subject of interpretability, James et al. [9] speak of "the trade-off between prediction accuracy and model interpretability" [p 24]. As statisticians and data scientists create models for predictive functions they may choose solutions that will generate accurate predictions at the expense of interpretability of results. These solutions (e.g. random forest) will typically use a very large number of data attributes (features) and even though they yield accurate predictions they have no explanatory power to help us understand the world and how it works.

On the subject of unintended consequences Wigan and Clark [10] highlight potential issues that could result from aggregating data generated from different sources, collected for different purposes, which could potentially lead to service denial to individuals. They mention that without reasonable explanation, "service denial has been increasingly apparent in many contexts, including government licensing, financial services, transport, and even health" [10, p 47].

Lack of privacy and the ability to control the usage of personal data are also important issues. For example, "analysts have documented examples of new kinds of inferences that can be drawn from this vast volume of data, along the lines of 'your social media service knows you're pregnant before your father does'" [10, p 49]

5. REFERENCES

[1] The data deluge | The Economist. Web Page, Feb. 2010. From printed edition www.economist.com/printedition/2010-02-27, The data deluge (accessed 10/27/2016).

[2] Mary Meeker. 2016 internet trends. http://www.kpcb.com/file/2016-internet-trendsreport. (accessed on 10/26/2016).

[3] Mary Meeker. 2014 internet trends. http://www.kpcb.com/blog/2014-internet-trends. (accessed on 10/27/2016).

[4] T. Davenport. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Publishing. Harvard Business Review Press, 2014.

[5] T. Hey, S. Tansley, K. M. Tolle, et al. The fourth paradigm: data-intensive scientific discovery, volume 1. Microsoft research Redmond, WA, 2009.

[6] Analytics: What it is and why it matters | SAS. http://www.sas.com/en us/insights/analytics/whatis-analytics.html. (accessed on 08/29/2016).

[7] Data, data everywhere | The Economist. http://www.economist.com/node/15557443. (accessed on 10/27/2016).

[8] Google now personalizes everyone's search results. http://searchengineland.com/google-now-personalizeseveryones-search-results-31195. (accessed on 08/29/2016).

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning, volume 6. Springer, 2013.

[10] M. R. Wigan and R. Clarke. Big data's big unintended consequences. Computer, 46(6):46–53, 2013.