“If we didn't solve small data in the past, how can we solve Big Data today?”

Purdue University

Spring 2018

John Springer-CNIT 559

Benya Chongolnee

Abstract

In the modern world, where every aspects of life is very technology-involved, there is constantly new data generated. Now, the world revolves around data that humans produce. The data produced can become useful when used correctly. This is where the terms such as small data and big data come in. This research paper will dive deep into the definitions as well as the benefits and challenges of small data and big data in order to answer the research question of “If we didn’t solve small data in the past, how can we solve Big Data today?” The purpose of this research paper is to better understand the importance of big data in the modern world. The paper will include topics such as “What is Small Data?”, “Why Small Data failed?” and “How Big Data Can Benefit Everyone”. After looking at resources from over ten different credible journals, textbooks, and articles, it can be understand that small data and big data go hands in hands. Small data alone cannot be solved to make predictions, but big data can be solved in order to create small data. This small data solved is the prediction to help businesses create insights to allow them to thrive. Simply, small data is unable to *create* insights, it *is* the insight gathered from big data.

Introduction

The weather is something that should be predicted correctly. If the forecast says that next Monday, Indiana will have a high of 71-degree Fahrenheit and low of 58-degree Fahrenheit, those that are visiting Indiana may bring only a light sweater with them on the trip. However, when Monday arrives, it turns out to be cold and rainy at the high of 45-degree Fahrenheit all day. This false prediction can ruin plans for many people that only arrived in Indiana with a light jacket. They may choose to either buy a new jacket or deal with the cold all day long. With that being said, weather and so many other variables need to be predicted accurately. One wrong prediction is able to turn a person’s life around as well.

A more extreme case of the wrong prediction is self-driving cars. According to Bettencourt (2014) from Santa Fe Institute, self-driving cars must “be able to measure and recognize potential problems just as they start to arise (car approaching) and act to make the necessary corrections (break or get out of the way, thereby avoiding collision)” (Bettencourt, 2014). If the cars fail to recognize problems, they may collide with another car; thus, killing lives all around. In this case, the failure to predict future problems is able to cost lives. This is when it is extremely crucial to accurately predict certain situations. In the past, there are many methods that could help makes the prediction more accurate. However, some have failed. The question that the modern world needs to address is

“If we didn't solve small data in the past, how can we solve Big Data today?”

What is Small Data?

Before diving straight into the big question, it is very important to first understand what small data is. Small data can have many different meanings all across the board. According to Kitchin (2017), data can be explained as elements that are extracted from observations, computations, experiments and record keeping. It is important to understand that the definition of data has changed over time since the seventeenth century. Over time, data had evolved into being "pre-analytical and pre-factual" which are fundamental elements of facts, evidence, information and knowledge (Kitchin, 2017).

Even though data is crucial to the modern world, it is expensive and difficult to store, analyze and/or process data. Because of this, data are produced using sampling techniques that limit the size and the scope of the actual data. This is called small data. Small data is proven to be successful in the past with a history of successful methodologies and modes of analysis (Kitchin, 2017). Small data are fixed data that does not allow much flexibility. For example, small data may include only information about countries and states; while big data would include information about individuals and households. Because of this, small data is able to answer specific research questions about ways people interact and make sense of the world. They are also less resource-intensive than big data. As described by Kitchin, small data and big data can easily be compared in Table 1.

	Small Data	**Big Data**
Volume	Limited-Large	Very large
Exhaustivity	Samples	Entire population

Resolution and Identification	Coarse and weak-tight and strong	Tight and strong
Relationality	Weak-strong	Strong
Velocity	Slow, freeze-framed/bundled	Fast, continuous
Variety	Limited-wide	Wide
Flexible and Scalable	Low-middle	High

Table 1: Small Data and Big Data comparison

Some people favor using small data over big data for many reasons. The first reason is that small data is “simpler and better suited to individual, everyday problems” (Huang, P & Huang, P, 2015). This is because the data that was collected can be applied individuals that the data is about only. As a consequence, small data can be referred as customized data. Another reason people prefer small data is that small is highly accurate. It is more accurate than big data because all the data collected were collected on purpose; thus, there is no unnecessary data in the dataset.

Why Small Data failed?

Since small data has many strong arguments as to why it is better than big data, people may ask why small data failed. Small data has its limitations in size and scope compared to big data. From Table 1, it can be seen that small data is not as specific as big data. Since small data is so limited, there is no room to expand and improve the dataset. As mentioned before, small data can answer many questions about the world; however, it is unable to dive more specific than that and expand on their analysis. More value and insight is able to extract from data if the datasets are bigger. Bigger datasets mean that the algorithms used to better understand the data and predict new data are more accurate. Consequently, small data failed because it is too simple and can only help individualized, customized problems. Small data is only able to answer specific research questions, but not predict new values or gain new insights. Small data is unable to apply its answers to a big group of people all at once. Furthermore, bigger data sets can be store to reuse or for future generations to take advantage of. Though gathering big data requires more funding, many organizations understand the importance and benefits of big data; thus, they are more willing to invest in big data knowing that it will be able to benefit from such investment (Kitchin, 2017).

What is Big Data?

Big data was previously discussed a bit in the previous section, but what exactly is big data? Just like small data, there are also many different meanings for big data. According to the textbook *Data Architecture: a Primer for the Data Scientist*, big data is data that is stored in very large volumes, stored on inexpensive storage, managed by the “Roman census” method, and stored and managed in an unstructured format (Linstedt & Inmon, 2015). However, an article from *Communications Today* stated that “big data is high-volume, high-velocity, and/or high-variety information that requires new forms of processing to enable enhanced decision making, insight discovery, and process optimization” ("Roadmap ahead," 2016). Though these two definitions are different, there is one part of it that are common, which is that big data has large volumes. This is the complete opposite of small data, where small data is relatively limited in size.

According to Alexandru Adrian (2013) from American University in Bucharest, Romania, big data includes photos, location, 3D models compare to the tradition data which includes personal files, finances, and documents (Adrian, 2013). Adrian expressed that the collection of big data can help companies in the modern world reveal patterns that can be used in order to improve the business. This is very attractive for companies who are looking to increase profit or better understand customers. Big Data can also be used to predict the future, which is important for many companies to see which direction they are heading. For example, a department store may want to look at customers’ time spent in the store, amount of money spent, and age in order to predict which customer may not like new items that the store is debating about selling. The goal of big data is nothing like small data. Big data focuses on the bigger picture and help others understand it better instead of answering one research question. It *solves* the information collected to reveal useful information.

Alguliyev and Abbaslı from Institute of Information Technology of Azerbaijan National Academy of Sciences 9 mentioned that the technologies used for big data analysis include “MPP (Massively Parallel Processing) analytical platform systems, Cloud Services, Hadoop and MapReduce and NoSQL data warehouse management systems” (Alguliyev, Gasimova & Abbasli, 2017). The technology chosen depends on the end goal of the project because they are all used to find a way to extract the information from the big data. The process of analyzing and extracting the essential information can be divided into four steps: data collection, integration, analysis, real-world application. These four steps can be used to help reduce the size of big data into a smaller and easier to interpret data set.

A new and useful design pattern that can be adopted on Hadoop or a relational database is called Data Lakes. Data Lakes are interesting because they supply raw data that users need for data exploration and discovery-oriented forms of advanced analytics. Philip Russom states that Data Lakes can "be a consolidation point for both new and traditional data, thereby enabling analytics correlations across all data" (Russom, 2017). Data Lake is an effective design pattern for capturing a wide range of data types at large scale. It is optimized for the quick processing of raw, detailed data; thus, it has a lot of great potential in the future.

With a large volume of data comes a lot of challenges and obstacles. The first and maybe most obvious challenge is how long it takes to analyze large datasets. Nrusimham Ammu and Mohd Irfanuddin (2013) stated: "when new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria" (Ammu & Irfanuddin, 2013). Each index structure can only support some classes of criteria. Doing this is challenging when an issue needs to be addressed as soon as possible. Other challenges do exist as well such as data acquisition, storage, processing, data transport/dissemination, archiving, data management/curation, and security. For example, since big data is so large, there are a lot of data that need to be stored for long-term use. Finding a place to store everything can be challenging and may cost more than it is worth. Another challenge is data transport. In cases when real-time data is needed, transporting analyzed data to the place where the data can be used by the people is very critical. This transportation may take longer than expected and it may be challenging to do so. Lastly, security is a challenge for everything in the technology world, including big data. Other than ensuring that the data remains confidential, it is important to ensure that the data is not compromised and that it remains correct and accurate ("Roadmap ahead," 2016). It is crucial to address the challenges in order to use big data to its full potential and avoid failure.

The reason small data failed is different to the reasoning behind how big data can fail. Despite its challenges, big data is incredibly useful to analyze and understand datasets. Once these challenges are solved, “the key will be finding a solution that is cost effective.” ("Roadmap ahead," 2016). According to Ammu and Irfanuddin (2013), the modern world “must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data” (Ammu & Irfanuddin, 2013). Together, the modern world will be able to address big data challenges and overcome them to ensure that big data is useful in helping the success of the world. As big data technologies become more complex, there will be a higher need for a highly skilled workforce to help deal with this. This means that there will be more need for people interested in big data to help the field grow; thus, creating more jobs in the workforce.

How Big Data Can Benefit Everyone

By understanding more about big data and its purpose, many opportunities will rise from the technology. Such opportunities include big data helping the government. Big data can help the government “harness and apply analytics to their big data” ("Roadmap ahead," 2016). The government in each country control a lot of every-day activities such as traffic control, land record, national security, and fraud. Big data can help manage these activities to better help the government and ensure that everyone is protected from other countries and each other. Other than protecting its citizens, big data can be used in election as well. For example, Barack Obama’s successful re-election was helped by big data analysis ("Roadmap ahead," 2016).

Another opportunities that big data can help is in communication. Big data is able to provide communication service providers information about location data generated by mobile devices ("Roadmap ahead," 2016). This information can help business partners deliver services that gear toward their customers. The idea of personalization based on customers action is very important because it makes marketing a lot more effective. It ensures that customers are receiving advertisements that the majority of people are interested in. If customers continuously see products they are interested in, they will buy the product; thus, increasing the company’s revenue. Big data can also help companies create new products that can be utilized or liked by more people, again, increasing their revenue.

Another possibility that rise with the help of big data is urbanization. As more cities grow, there is a higher need for big data analytics. According to Bettencourt (2014), urban transportation systems should track the time in between buses at each stop, number of passengers waiting, and more to find out if they need to add more buses to the system (Bettencourt, 2014). This is called feedback control theory, which provides the framework to help development cities optimize their cost while improving their citizens’ overall happiness. This theory can be used for so many other different elements within a city such as water, power supply management, traffic management, trash collection, and infrastructure maintenance. Overall, big data can help with so different aspects of life; thus, it can benefit many people all around.

The Question

“If we didn't solve small data in the past, how can we solve Big Data today?”

As stated previously, small data are fixed data that does not allow much flexibility, but it is still useful to answer specific research questions. Small data simply cannot be solved to reveal patterns or insights. While large data are information that requires solving to enhance decision making, insight discovery, and process optimization. As stated by Po-Chieh and Po-Sen Huang, big data and small data are very similar because big data is solved to become small data later on. This small data is then can be utilized by the world. What do they mean by this? The example they gave was in regards to Bing Travel. Bing Travel used big data to predicts airfares in the future. This prediction, the small data, gives big data a real-world value (Huang, P. & Huang, P., 2015). Though small data alone cannot be solved to make predictions, big data can be solved in order to create small data. Po-Chieh and Po-Sen Huang believe that small data and big data go hands in hands. Small data gives big data a value and vice versa. Big data aggregates data sets, analyze them and look for meaningful patterns. Again, this pattern creates insights to help businesses grow. This is the goal of big data, not small data. Small data is unable to *create* insights, it *is* the insight.

Conclusion

Big data is a term that most people have heard, but do not quite understand. By understanding what big data is and how it can transform the world for the better, many will be able to recognize the significance of it. Big data can help solve problems in every aspect of the world: from retail to government. John Forsyth and Leah Boucher (2015) understand how big data provides actual “behaviour to quantify areas of growth” (Forsyth & Boucher, 2015). However, they argue that big data is unable to reveal “what consumers are thinking or why they behave the way they do” (Forsyth & Boucher, 2015). This is where market research should be integrated with big data to provide thorough insights that truly help companies understand its consumers. Big data alone will not be able to fully understand consumers’ behavior and predict if the predicted insight is truly accurate. Though there are challenges within big data that need to be addressed, big data can still reveal patterns that are able to give helpful insights. Ultimately, big data is a growing field that needs more research to reach its full potential. With that being said, big data still has a lot of great potential in the future.

References

Adrian, A. (2013). Big Data Challenges. *Database Systems Journal, IV*(3), 31-40. Retrieved from http://www.dbjournal.ro/archive/13/13_4.pdf

Alguliyev, R. M., Gasimova, R. T., & Abbasli, R. N. (2017). The obstacles in big data process. *International Journal of Modern Education and Computer Science, 9*(3), 28-n/a. Retrieved from https://search-proquest-com.ezproxy.lib.purdue.edu/docview/1886771939?accountid=13360

Ammu, N., & Irfanuddin, M. (2013). Big Data Challenges. *International Journal of Advanced Trends in Computer Science and Engineering,, 2*(1), 613-615. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.2518&rep=rep1&type=pdf

Bettencourt, L. (2014). The Uses of Big Data in Cities. *Big Data., 2*(1), 12-22.

Forsyth, J. and Boucher, L. (2015), Why Big Data Is Not Enough. Research World, 2015: 26-27. doi:10.1002/rwm3.20187

Huang, P., & Huang, P. (2015). WHEN BIG DATA GETS SMALL. *International Journal of Organizational Innovation (Online), 8*(2), 100-117. Retrieved from https://search-proquest-com.ezproxy.lib.purdue.edu/docview/1721368656?accountid=13360

Kitchin, R. (2017). *The data revolution: Big data, open data, data infrastructures & their consequences*. Los Angeles: Sage.

Linstedt, W. H., & Inmon, W. H. (2015). *Data Architecture: A Primer for the Data Scientist*. Morgan Kaufmann.

Roadmap ahead: Big data - big opportunities and big challenges. (2016). *Communications Today,* Retrieved from https://search-proquest-com.ezproxy.lib.purdue.edu/docview/1772876201?accountid=13360

Russom, P. (2017, March 29). Executive Summary | Data Lakes: Purposes, Practices, Patterns, and Platforms. Retrieved from https://tdwi.org/articles/2017/03/29/executive-summary-data-lakes.aspx